Understanding hidden neuron activations using structured background knowledge and deductive reasoning

by

Abhilekha Dalal

B.Tech, Amity University, 2017

M.S., Kansas State University, 2021

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

A central challenge in Explainable AI (XAI) is accurately interpreting hidden neuron activations in deep neural networks (DNNs). Accurate interpretations help demystify the black-box nature of deep learning models by explaining what the system internally detects as relevant in the input. While some existing methods show that hidden neuron activations can be human-interpretable, systematic and automated approaches leveraging background knowledge remain underexplored. This thesis introduces a novel model-agnostic post-hoc XAI method that integrates a Wikipedia-derived concept hierarchy of approximately 2 million classes as background knowledge and employs OWL-reasoning-based Concept Induction to generate explanations. Our approach automatically assigns meaningful class expressions to neurons in the dense layers of Convolutional Neural Networks, outperforming prior methods both quantitatively and qualitatively.

In addition, we argue that understanding neuron behavior requires not only identifying what activates a neuron (recall) but also examining its precision—how it responds to other stimuli, which we define as the neuron's error margin, enhancing the granularity of neuron interpretation.

To visualize these findings, we present ConceptLens, an innovative tool that visualizes neuron activations and error margins. ConceptLens offers insights into the confidence levels of neuron activations and enables an intuitive understanding of neuron behavior through visual bar charts. Together, these contributions offer a holistic approach to interpreting DNNs, advancing the explainability and transparency of AI models.

Understanding hidden neuron activations using structured background knowledge and deductive reasoning

by

Abhilekha Dalal

B.Tech, Amity University, 2017

M.S., Kansas State University, 2021

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Pascal Hitzler

Copyright

Abstract

A central challenge in Explainable AI (XAI) is accurately interpreting hidden neuron activations in deep neural networks (DNNs). Accurate interpretations help demystify the black-box nature of deep learning models by explaining what the system internally detects as relevant in the input. While some existing methods show that hidden neuron activations can be human-interpretable, systematic and automated approaches leveraging background knowledge remain underexplored. This thesis introduces a novel model-agnostic post-hoc XAI method that integrates a Wikipedia-derived concept hierarchy of approximately 2 million classes as background knowledge and employs OWL-reasoning-based Concept Induction to generate explanations. Our approach automatically assigns meaningful class expressions to neurons in the dense layers of Convolutional Neural Networks, outperforming prior methods both quantitatively and qualitatively.

In addition, we argue that understanding neuron behavior requires not only identifying what activates a neuron (recall) but also examining its precision—how it responds to other stimuli, which we define as the neuron's error margin, enhancing the granularity of neuron interpretation.

To visualize these findings, we present ConceptLens, an innovative tool that visualizes neuron activations and error margins. ConceptLens offers insights into the confidence levels of neuron activations and enables an intuitive understanding of neuron behavior through visual bar charts. Together, these contributions offer a holistic approach to interpreting DNNs, advancing the explainability and transparency of AI models.

Table of Contents

List of Figures

Acknowledgments

This journey would not have been possible without the support and encouragement of many incredible individuals. First and foremost, I am deeply grateful to my advisor, Pascal Hitzler, for his invaluable guidance, patience, and unwavering belief in my abilities. His mentorship has been instrumental in shaping both my research and my growth as a scholar.

I would also like to extend my heartfelt thanks to my lab-mates and collaborators, who inspired me to think beyond my comfort zone and explore new perspectives. Working alongside you has been an enriching experience that has greatly broadened my understanding of the field.

To my friends, thank you for standing by me, even during times when I was entirely consumed by my work. Your support means the world to me, and I am immensely grateful for all the moments of joy and laughter that kept me grounded.

Lastly, I want to thank my family for their endless love and encouragement. Your belief in me has been my anchor, and I could not have come this far without you.

The authors acknowledge partial funding under the National Science Foundation grants 2119753 "RII Track-2 FEC: BioWRAP (Bioplastics With Regenerative Agricultural Properties): Spray-on bioplastics with growth synchronous decomposition and water, nutrient, and agrochemical management" and 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."

Dedication

To my parents, advisor, and friends who believed in me and are eagerly waiting for me to become Dr. Abhilekha Dalal.

Chapter 1

Introduction

The increasing prominence of artificial intelligence (AI) has led to the widespread adoption of deep neural networks across a diverse range of real-world applications. Neural networks have displayed exceptional performance, rivaling human capabilities in areas such as image classification[1], speech recognition[2], and translation[3]. They have even demonstrated mastery in complex tasks like strategic gameplay[4], and have been deployed in critical fields including medical diagnosis[5], drug discovery[6], and autonomous vehicles[7]. These successes, however, have highlighted a major limitation: the "black box" nature of deep learning models, which renders their decision-making processes opaque.

This opacity poses significant challenges in domains where trust, transparency, and accountability are paramount, such as safety-critical systems and automated decision-making processes. For instance, biases in AI have led to real-world controversies like Apple co-founder Steve Wozniak accused Apple of gender discrimination, claiming that the new Apple Card gave him a credit limit that was ten times higher than that of his wife even though the couple shares all property[8]. In an image search, only 11% of the top image results for "CEOs" were images of women despite the fact that women make up 27% of US CEOs[9]. In another study, CNNs trained on non-biomedical image segments were still able to classify biomedical images with high accuracy, raising concerns about the reliance of models on irrelevant features or data artifacts[10]. These challenges are particularly critical in areas such as self-driving

cars[11], drug discovery, and treatment recommendations[12;13], where deep learning systems are vulnerable to adversarial attacks that can manipulate results by introducing adversarial examples or even by controlling the order of training data[14;15]. Some of these attacks are difficult to detect after the fact, as represented in[16;17]. These examples underscore the need for more interpretable AI systems, especially in applications where biases or errors can have significant consequences.

While deep learning models achieve remarkable results, their lack of transparency in neural network models can be attributed to their considerable size and complexity. These models often encompass a huge number of parameters, sometimes reaching into the billions, distributed across numerous layers, each engaged in complex non-linear transformations[18]. Standard assessments of deep learning performance consist of statistical evaluation but do not seem sufficient to address these shortcomings as they cannot provide reasons or explanations for particular system behaviors[19]. However, neural networks are subsymbolic in nature, implying that their internal hidden layer representations rely on high-dimensional Euclidean spaces, such as real-valued vectors and matrices, which lack any associated declarative meaning[20] as a result, understanding their inner workings becomes even more challenging.

The opacity of these models piqued researchers' interest, prompting them to investigate the underlying mechanisms behind their impressive results. This investigation gave rise to the field of explainable AI (XAI), which focuses on developing techniques to make AI systems more comprehensible to human users[21]. XAI research aims to uncover the internal processes of deep learning models, making it easier to trust, debug, and deploy them in critical environments. Despite notable progress in Explainable AI (XAI), several critical gaps persist in current approaches. Traditional methods often rely on a limited set of hand-picked explanation categories curated by domain experts. While useful, these predefined categories are inherently restrictive, assuming that they are universally applicable across various tasks. This assumption remains unsubstantiated, particularly in specialized fields requiring more nuanced insights. Many XAI techniques also depend on intricate low-level data features projected into higher-dimensional spaces, making them accessible only to domain experts[22–24]. Additionally, these methods have shown vulnerability to adversarial tampering, where alter-

ing attributed features may not impact the model's decisions, further reducing trust in these systems[25–27].

State-of-the-art methods that modify deep learning architectures to enhance explainability often lead to performance degradation compared to unmodified models[28]. Similarly, approaches using deep learning models like Large Language Models (LLMs) to generate explanations merely shift the "black box" problem onto the explanation generation itself[29]. To truly advance the integration of deep learning into safety-critical systems, an ideal XAI method should offer self-explanatory, model-agnostic solutions that provide human-understandable explanations without compromising performance or accuracy.

Concept-based XAI methods offer a promising solution by linking high-level, human-understandable concepts to the activations of hidden neurons in deep neural networks. However, these methods also have significant limitations. Many approaches rely on manually selected concepts measured for their correlation with model outcomes. A major issue remains unresolved: can a limited set of chosen concepts provide a comprehensive understanding of the model's decision-making process ? Some implementations[29] often select the most frequent English words as candidate concepts, which may suffice for general tasks but lack the specificity required for specialized domains such as healthcare or genomics. These gaps underscore the need for more scalable, robust XAI methods that offer concept-rich, interpretable explanations without compromising the reliability or performance of the model, particularly in domains where trust and transparency are paramount.

This thesis introduces a novel framework for enhancing the explainability of deep learning models by leveraging *Concept Induction* i.e., formal logical deductive reasoning[30] . Grounded in formal logic, this approach utilizes structured knowledge bases, such as Wikipedia[31] (which has class hierarchy consisting of about $2 \cdot 10^6$ classes) as the pool of categories, to automatically generate comprehensive, interpretable explanations for hidden neuron activations in a Convolutional Neural Network (CNN) architecture for image scene classification (on the ADE20K dataset[32]). Unlike traditional methods that rely on predefined explanation categories, this framework dynamically constructs explanations from a vast pool of interconnected concepts, offering greater flexibility and granularity.

While associating neurons with high-level, human-understandable concepts shows promise in explaining deep learning models, a major challenge lies in the precision of these concept labels. Neurons often exhibit distributed activations, responding not only to inputs tied to their assigned labels but also to unrelated stimuli. This multi-response behavior leads to a high false-positive rate, where neurons activate for inputs that don't match their intended concept label. For example, a neuron labeled as detecting "cat" may also activate in response to stimuli like "dog" or "tiger." This ambiguity reduces the interpretability and reliability of the explanations, as users may be misled into believing a neuron represents a specific concept when it does not. If the system cannot guarantee that a neuron's activation corresponds consistently to its intended concept, it becomes difficult to reliably interpret the network's true detections, especially in applications requiring clear and accurate interpretations of the model's decision-making processes such as in safety and healthcare. This issue diminishes the model's practical utility, particularly in contexts requiring transparent and trustworthy AI decision-making.

To address this, the thesis introduces an error-margin analysis that assigns confidence levels to neuron activations, improving the precision and reliability of the generated explanations. This statistical framework quantifies the likelihood that a neuron activation corresponds to its assigned concept label, significantly reducing false positives and enhancing the trustworthiness of concept-based explanations.

Our approach is driven by several key principles: explanations should be easily understood by end-users without deep learning expertise, and there should be an organized structure of human-understandable concepts with clear relationships. Relevant concepts for model explanations should be automatically extracted to avoid manual bias, the explanation technique should be interpretable, avoiding black-box methods, and should be rigorously evaluated across multiple dimensions. Finally, explanations should include assigned confidence levels to improve the precision and reliability of explanations by reducing false positives. The core contributions of this thesis are:

1. A zero-shot, model-agnostic XAI method that interprets hidden neuron activations in

pre-trained models using high-level concepts from knowledge graphs. It is inherently interpretable, scalable, and achieves state-of-the-art performance without modifying model architecture.

2. A comparative analysis demonstrating that the proposed method outperforms leading XAI techniques, such as CLIP-Dissect[29] and GPT-4[33].

3. A fully automated end-to-end system to use Concept Induction to interpret neurons' activations in terms of concepts along with evaluation in a pre-trained model(CNN).

4. An error-margin analysis that quantifies the precision of concept-based explanations, improving neuron label accuracy and reducing false positives.

5. ConceptLens: A tool designed to represent concepts with a bar chart displaying confidence measures for neuron activations in a CNN.

By addressing the challenges, our research advances the field of Explainable AI, providing a robust and interpretable framework for real-world applications. In safety-critical systems, where understanding AI decision-making is essential, the combination of Concept Induction and error-margin analysis lays the foundation for developing more transparent, trustworthy, and actionable AI systems. The benefits of our explainable deep learning approach include: (a) compatibility with unmodified and pre-trained architectures, (b) the assignment of semantic categories (class labels expressed in formal logic) to hidden neurons, ensuring that images related to these labels activate the corresponding neuron with high probability, and (c) the ability to construct these labels from a vast pool of categories.

1.1 Outline

This thesis is a cumulative dissertation that focuses on enhancing the explainability of deep neural networks through innovative approaches that integrate symbolic reasoning. As mentioned in the above introduction, methods introduced in this work aim to provide transparent and interpretable AI systems that can better communicate their decision-making processes.

This work can be divided into two concrete research topics that incrementally build towards this goal.

The remainder of this dissertation is outlined as follows:

Chapter 2 provides a comprehensive background on the essential concepts related to the thesis, including the architecture of Convolutional Neural Networks (CNN), the role of knowledge graphs, and the process of Concept Induction. This chapter establishes the foundational principles necessary for understanding the subsequent research contributions. Chapter 3 reviews literature on explainable AI, concept-based neuron labeling, and neural network interpretability. It outlines key advancements, identifies gaps, and contextualizes the research. This review establishes the foundation for the novel contributions in subsequent chapters by highlighting existing limitations and motivating new approaches.

Chapter 4 presents the first research topic: providing meaningful interpretations for hidden neuron activations in Convolutional Neural Networks (CNNs). It introduces the key 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? The chapter discusses the use of a Wikipedia-derived concept hierarchy and deductive reasoning-based Concept Induction to offer explanations. Additionally, it evaluates the proposed neurosymbolic method against other explainable methods, comparing its quantitative and qualitative performance. The primary contributions referenced in this section are:

- *On the Value of Labeled Data and Symbolic Methods for Hidden Neuron Activation Analysis*[34]
- *Automating CNN Neuron Interpretation using Concept Induction*[35]

Chapter 5 presents the second research topic: introducing error margins for the concept of neuron labels to enhance the interpretability of neuron activations in Convolutional Neural Networks (CNNs). It puts forward the key research question: how can we assess and improve the precision of neuron labels in order to minimize false-positive activations? This chapter discusses the extension of prior concept-based neuron labeling techniques by introducing

error margins for neuron activations. It also provides statistical validation of the error margin approach through user experiments, highlighting its impact on improving explainability in deep learning models. The primary contributions referenced in this section are:

- *Error-margin Analysis for Hidden Neuron Activation Labels*[36]
- *ConceptLens: from Pixels to Understanding*[37]

Chapter 6 presents concluding remarks through a brief summary that highlights the overall contributions and how they fit together, especially with respect to the state of the art. Additionally, I provide an outlook on future work.

The appendices contain the full publications that form the foundation of this dissertation, listed in order of appearance:

- Appendix A: Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y. Vasserman, Md Kamruzza- man Sarker, and Pascal Hitzler. *On the Value of Labeled Data and Symbolic Methods for Hidden Neuron Activation Analysis*. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, Neural-Symbolic Learning and Reasoning, pages 109–131, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.)
- Appendix B: Samatha Ereshi Akkamahadevi, Abhilekha Dalal, and Pascal Hitzler. *Automating CNN Neuron Interpretation using Concept Induction*. In International Semantic Web Conference, 2024. To appear.
- Appendix C: Abhilekha Dalal, Rushrukh Rayan, and Pascal Hitzler. *Error-margin Analysis for Hidden Neuron Activation Labels*. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, Neural-Symbolic Learning and Reasoning, pages 149–164, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.
- Appendix D Abhilekha Dalal and Pascal Hitzler. *Conceptlens: from pixels to understanding*, 2024. URL https://arxiv.org/abs/2410.05311.

Chapter 2

Background

2.1 Convolutional Neural Networks (CNNs)

CNNs are specifically designed to process and analyze visual data, making them exceptionally well-suited for tasks like image recognition, object detection, and image classification.

Convolutional neural networks (CNNs) can be expressed mathematically as a composition of linear and non-linear tensor maps called layers. A tensor is a multidimensional array of real numbers, and the input to a CNN is a tensor representing an image. The initial convolutional layer applies learnable filters to generate feature maps that capture spatial information. Subsequently, an activation function introduces non-linearity, and these feature maps are downsized by pooling layers, preserving essential information. Following this are fully connected layers, where neurons are connected to each other and are used to learn high-level features from the feature maps produced by the convolutional and pooling layers. The output of the last fully connected layer is a tensor representing the probability that the input image belongs to each class. This tensor can then be used to classify the image. Figure 2.1 shows a CNN architecture with three convolutional layers, each accompanied by an activation function and pooling layer, followed by a fully connected layer that produces class probabilities for the input.

Training a CNN includes minimizing a loss function through optimization algorithms like

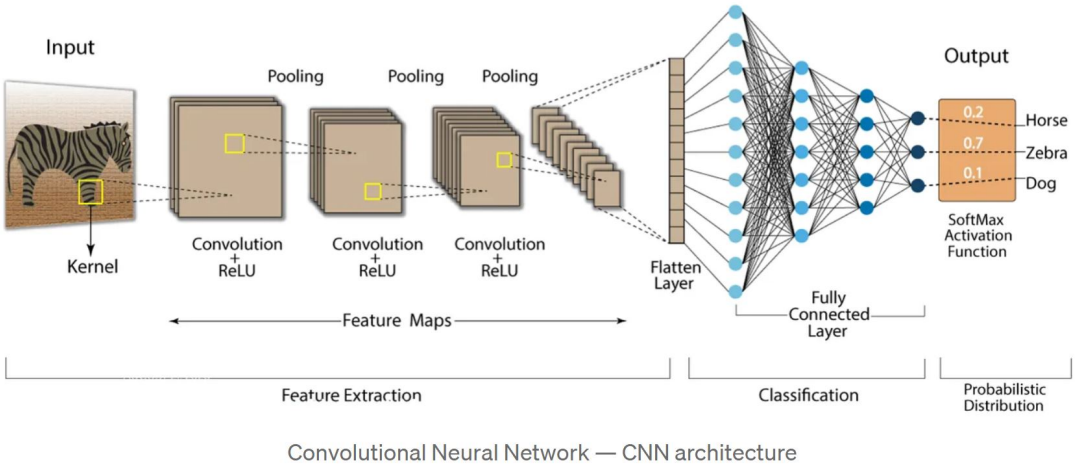

Figure 2.1: *Simple architecture of convolutional neural network*[38].

gradient descent. This process adjusts filter and neuron weights and biases to reduce the gap between predicted outputs and actual labels, a technique called backpropagation. It iteratively refines network parameters, enhancing performance over time.

2.2 Knowledge Graphs

Knowledge Graphs are powerful data structures designed to represent entities and their relationships in a structured, machine-readable format. They are typically represented as directed, edge-labeled graphs, where nodes represent entities (or objects) and edges represent relationships between them. These relationships are often defined within an ontological schema, providing a formal structure that allows for logical reasoning over the data.

A key feature of knowledge graphs is their organization into two components:

- TBox (Terminology Box): This refers to the schema or ontology, defining the concepts (classes) and relationships (properties) that exist in the graph. It lays the foundation for how data is structured and understood.
- ABox (Assertion Box): This contains the actual data—individuals (instances) and their relationships—asserting facts based on the structure provided by the TBox.

Formally defining let C denotes a set of entities (also known as constants) and P for

predicates, KG denoted as $G \subseteq C \times P \times C$ signifies a collection of factual information stored in triplets in the format of (*subject, predicate, and object*). For example, a simple fact like example (*street, has, cross_walk*) expresses a relationship between a street and a crosswalk, where "has" is the predicate linking the two.

Knowledge Graphs can be referred to as Linked Data[1], which is a method for interconnecting structured data on the web in a meaningful way. The Linked Data paradigm encourages data to be shared in a standardized, open format, allowing information from various sources to be connected and queried seamlessly. Reasoning over these knowledge graphs is achievable through standard knowledge representation formalisms such as RDF[2], RDFS[3], and OWL[4]. These formalisms enable reasoning over the knowledge graph, meaning that implicit facts (inferred relationships) can be derived from explicit information using logical rules. For example, if a knowledge graph contains the facts *(car,is_a,vehicle)* and *(vehicle,has,wheels)*, it may infer that *(car,has,wheels)*.

Knowledge graphs have evolved into intricate structures resulting from the integration, extraction, and manipulation of data from diverse sources on a massive scale. Various knowledge graphs are now accessible on the web, including examples like DBpedia[5], WikiData[6]. In this proposal, our primary focus is on leveraging the knowledge graph derived from Wikipedia[31].

2.3 Concept Induction

Concept Induction[30] is a technique used to automatically generate meaningful, human-understandable labels or descriptions for sets of data points based on a given knowledge base, typically an ontology or a knowledge graph. The process relies on deductive reasoning over description logics—a formal framework used to represent the structure and semantics of

[1]https://www.w3 .org /wiki /LinkedData.
[2]https://www.w3.org/RDF/
[3]https://www.w3.org/wiki/RDFS
[4]https://www.w3.org/OWL/
[5]https://www.dbpedia.org.
[6]https://www.wikidata.org /wiki /Wikidata:Main _Page.

knowledge graphs, ontologies, and the broader Semantic Web[39;40].This technique has been widely used in ontology engineering to derive new knowledge or (TBox) axioms from existing base data (ABox) graph, enhancing the structure and utility of ontologies. Concept Induction has indeed already been shown, in other scenarios, to be capable of producing labels that are meaningful for humans inspecting the data[41].

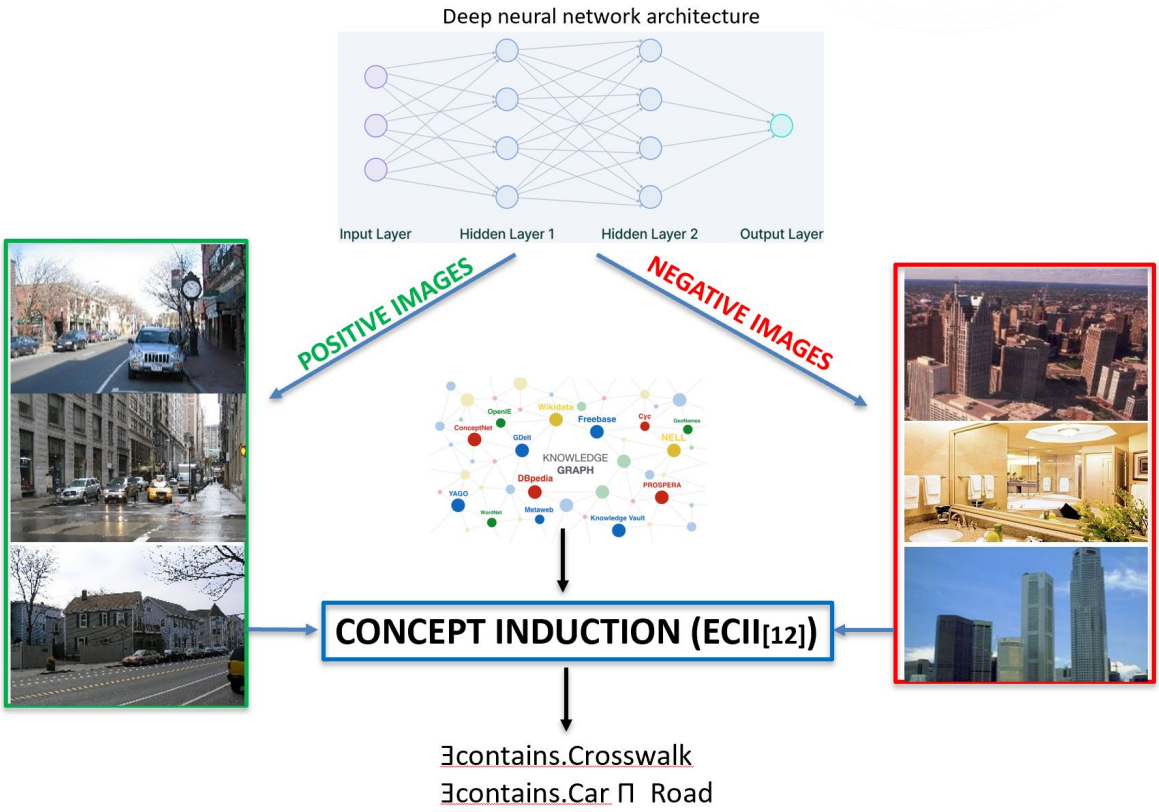

Figure 2.2: *The figure illustrates the process of Concept Induction used to interpret hidden neuron activations in a deep neural network. Positive images (left) and negative images (right) are selected based on neuron activation patterns. A knowledge graph is employed to perform semantic reasoning and Concept Induction, generating interpretable concepts.*

A Concept Induction system accepts three inputs: (1) a set of positive examples P, (2) a set of negative examples N, and (3) a knowledge base (or ontology) K, all expressed as description logic theories, and all examples $x \in P \cup N$ occur as instances (constants) in K. It returns description logic class expressions E such that all the positive examples are contained in each of the class expressions while none of the negative examples is, represented as $K \models E(p)$ for all $p \in P$ and $K \not\models E(q)$ for all $q \in N$. If no such class expressions exist,

then it returns approximations for E together with a number of accuracy measures.

In simpler terms, Concept Induction helps to create new logical concepts or class expressions by analyzing examples of what belongs to a certain category and what does not. The generated concepts are meaningful and interpretable to humans, making it a valuable tool for tasks requiring explanation or reasoning over large datasets. For the purpose of this thesis, we will be using ECII[42] as the Concept Induction system.

Chapter 3

Related Work

Explaining (interpreting, understanding, justifying) automated AI decisions has been explored from the early 1970s. With the recent advances in deep learning[43], 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[44–46]. In this chapter, a comprehensive review of the current state-of-the-art literature on Explainable AI (XAI) techniques applied to deep learning models is presented, with focus on explaining Convolutional Neural Networks (CNNs). Section 3.1 and 3.2 introduce existing explanations of feature attribution-based methods, which can be classified into two primary categories: local explanations and global explanations. Section 3.3 presents approaches integrating external domain knowledge, i.e., knowledge graphs with deep representations to generate semantic explanations.

3.1 Local Explanations

Local explanation algorithms aim to understand why a particular prediction was made. A common approach for achieving local explanations is to utilize visualization techniques such as saliency maps[47;48] to highlight the parts of an image that played the most significant role in the model's prediction. Many studies and surveys have explored visualization techniques[49;50] like backpropagation-based methods[51], where a set of propagation rules is used to visualize

feature relevance by tracking the volume of gradients passed through layers during network training. Perturbation-based approaches[52] involves visualizing feature relevance by comparing the network output of an input when different parts of the input are masked. A more flexible way to analyze the outcome of a DNN locally is by using prototype/example-based algorithms such as influence functions[53], which have been used to explain model decisions by referencing other instances (prototypes or examples) similar to the input in question. Other methods include proxy-based methods like LIME[54], which aim to mimic the behavior of complex models by constructing a simpler and interpretable model (proxy model often black-box) that can be easily explained for a small set of input instances. Another kind of approach uses a phrase-critic model for extracting both factual and counterfactual explanations for image classification models[55].

While the above-mentioned methods increase the interpretability of neural network models in some way, majority of these methods do so in terms of the model's input by generating explanations consisting only of sets of input features. Such explanations require human interference to interpret each case or when dealing with highly unstructured data, which can hinder scalability and introduce the potential for human bias to influence the interpretation process. Additionally, certain methods, including proxy-based approaches, may introduce an additional layer of complexity and opacity in the explainability process.

3.2 Global Explanations

Global explanation algorithms are designed to provide a comprehensive understanding of the overall behavior of deep learning models. One common approach is to use visualization techniques that can help understand the functions of convolutional filters or entire convolutional layers by examining their visual representations[56]. Another strategy is based on rule extraction, which aims to extract rules that explain the decisions made by complex models. Rule extraction methods can be broadly categorized into three main types: pedagogical rule extraction algorithms extract rules from neural networks solely based on their input-output behavior without delving into the model's internal architecture[57], decompositional

rule extraction algorithms derive rules by analyzing the activations and weights within the inner structure of neural network models[58], eclectic rule extraction algorithms are a hybrid approach that uses elements of both pedagogical and decompositional rule extraction algorithms[59]. Explanations can also be provided in terms of semantic concepts, and one such method is TCAV (Testing with Concept Activation Vector)[60], which is developed to quantitatively assess the significance of specific semantic concepts on the prediction of output classes in classification problems. It works by defining set of concept activation vectors (CAVs), which are representations of the input features, and then uses directional derivatives to quantify the sensitivity of the model's prediction to each CAV.

Nonetheless, these methods face limitations; either they are not suitable for explaining convolutional layers, or they struggle to explain the inner workings of hidden layers. In cases where they are applied to layers, like in visual analyses, they often rely on human interpretation, thereby hindering scalability. Even methods like TCAV, which are good at interpreting results using semantic concepts, struggle when concepts are related, which is common in real-world image datasets[61]. Furthermore, none of the methods mentioned above is able to leverage external knowledge bases to generate explanations.

3.3 XAI Approaches using Knowledge Graphs

Prior work has shown that internal units may indeed represent human-understandable concepts[14;62], but these approaches require semantic segmentation[63] Justifying the result of a neural network requires a defined language that incorporates elements of reasoning that use knowledge bases to create human-understandable yet unbiased explanations. Knowledge bases such as DBPedia or Wikidata can serve as effective tools for explaining deep neural networks, especially when the aim is to produce self-explanatory and directly understandable explanations in human terms. There has been some work using knowledge graphs to produce explanations from deep learning models[64;65], and also on using Concept Induction to provide explanations[66;67] where the authors use the Suggested Upper Merged Ontology (SUMO) along with a DL-learner to generate explanations for image classifiers, but they focused on

analysis of input-output behavior, i.e., on generating an explanation for the overall system. This proposal focuses instead on the different task of understanding internal (hidden) node activations.

The work most similar to ours similar in spirit demonstrated on a similar use case, but using a completely different approach, comes from[29] who use the CLIP pre-trained model, employing zero-shot learning to associate images with labels. Label-Free Concept Bottleneck Models[68] builds upon CLIP-Dissect and uses ChatGPT for concept set generation. However, CLIP-Dissect as currently presented, exhibits certain limitations like generating low-level concepts when needed (e.g., identifying "Zebra" upon observing stripes) and challenges in transferring the approach to different modalities or domain-specific applications due to the absence of an off-the-shelf concept set and domain-specific pre-trained models. Furthermore, Label-Free Concept Bottleneck Models inherit the foundational limitations of CLIP-Dissect and introduce potential issues with explainability by using a concept derivation method that lacks inherent explainability.

To the best of our knowledge, our proposed approach of utilizing Concept Induction with extensive background knowledge as a resource for generating explanations, specifically for understanding internal node activations, is a novel contribution. Concept Induction not only provides access to background concepts but also allows for the generation of conjunctions and a diverse list of concept outputs to choose from, catering to different levels of abstraction in concept understanding and is not dependent on the availability of a pre-existing concept set and a domain-specific pre-trained concept generation model. Our proposed methodology (training, Concept Induction analysis, and verification) is flexible and fully automatable eliminating the need for human intervention. This approach can be applied effectively to pre-trained models as well as models trained from scratch. Most notably, our work has the unique ability to generalize across various neural network architectures, making it adaptable and applicable to various scenarios.

Chapter 4

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

4.1 Overview

Deep learning models have achieved remarkable performance across various domains, yet their decision-making processes often remain opaque to human understanding. This lack of transparency poses significant challenges, particularly in critical applications where explanations are crucial for trust and accountability. Most current methods to improve explainability focus on deriving explanations purely from the feature space used during training, relying on predefined explanation categories. These categories, however, often lack human intuitiveness and fail to incorporate domain knowledge, leading to explanations that are not easily understood by end-users.

Humans excel at grasping and manipulating abstract concepts. By framing explanations in terms of familiar ideas and their relationships, we can bridge the gap between complex mathematical operations and human-interpretable logic. There is growing recognition in the academic literature of the complementary strengths of subsymbolic AI—such

as deep learning—and symbolic AI—such as knowledge representation[69]. This synergy has driven increased interest in fusing data-driven neural methods with knowledge-driven symbolic approaches, creating opportunities to enhance the explainability, transparency, and trustworthiness of AI systems.

Knowledge graphs offer a structured representation of real-world information, capturing concepts and their interconnections. By integrating these rich semantic structures with deep learning models, we can map hidden neuron activations to meaningful concepts, apply deductive reasoning to infer new, implicit relationships, and generate more nuanced and contextually relevant explanations. Leveraging symbolic methods like knowledge graphs allows us to structure learned knowledge in ways that align more closely with human cognition, offering clearer, more intuitive insights into a model's decision-making process.

Previous research has shown promise in using semantic annotations to label internal components of neural networks[64;65]. However, these approaches often rely on simplistic domain knowledge, limiting the depth and sophistication of the resulting explanations[66;67]. Our work aims to address these limitations by generating human-understandable explanations for model decisions through an automated process that interprets neural network activations. We develop an approach that harnesses more expressive and comprehensive knowledge graphs to perform deductive reasoning over complex knowledge structures. This method enables us present conceptual explanations in an intuitive, human-friendly manner without manual intervention or sacrificing model performance. By bridging the gap between opaque neural computations and interpretable reasoning, our approach seeks to enhance model transparency while maintaining the high accuracy that deep learning systems are known for. For this topic, I formulated the following research question considering these concerns and are addressed in the next section.

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?

4.2 Contributions

[34] Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y. Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. *On the value of labeled data and symbolic methods for hidden neuron activation analysis.* In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, Neural-Symbolic Learning and Reasoning, pages 109–131, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.

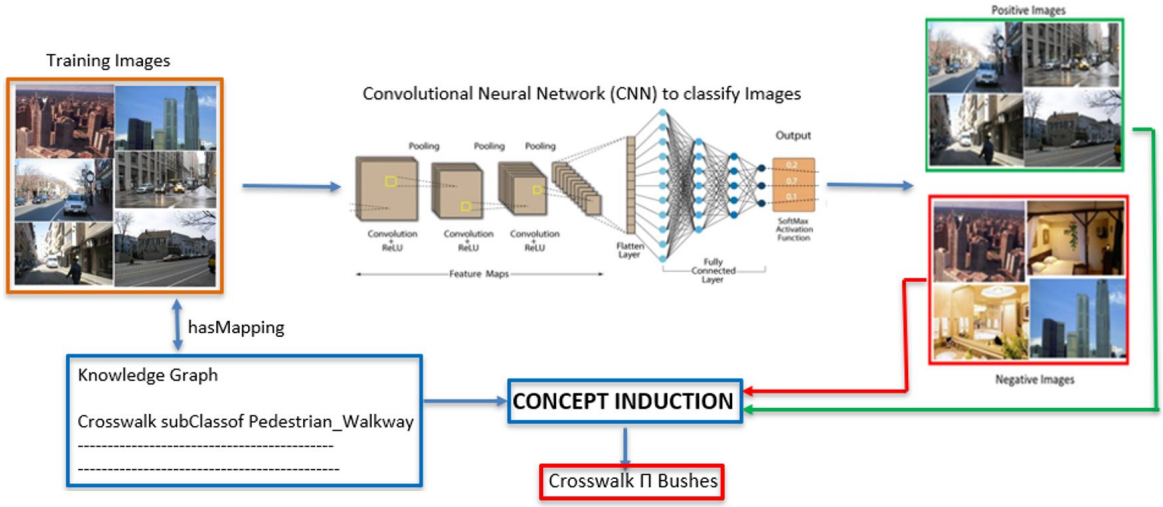

Figure 4.1: *The diagram shows the pipeline for interpreting hidden neuron activations utilizing CI in CNNs. Training images are used by CNN to classify positive and negative examples. A knowledge graph maps concepts, and Concept Induction is used to derive interpretable concepts.*

In the publication *On the value of labeled data and symbolic methods for hidden neuron activation analysis*[34], we address the research question of developing an effective approach to explainable deep learning by introducing a novel, model-agnostic Explainable AI method. This method provides meaningful, human-understandable interpretations of hidden neuron activations in the dense layer of Convolutional Neural Networks (CNNs) by leveraging deductive reasoning over a vast Wikipedia-derived concept hierarchy comprising approximately two million classes. By integrating Concept Induction system - ECII[42] with deep learning

models, our approach bridges the gap between abstract neural representations and human-intuitive concepts. It automatically extracts relevant concepts from the extensive background knowledge base, eliminating the need for manually selected explanation categories and ensuring a comprehensive, unbiased set of potential explanations. Concept Induction, based on formal logic reasoning, offers inherent explainability in the extraction and association of concepts with neuron activations, providing a transparent alternative to black-box models or opaque mechanisms.

To demonstrate the effectiveness of our method, we applied it to a pre-trained CNN architecture (ResNet50V2) designed for image scene classification, utilizing the ADE20K dataset, which contains over 27,000 images across 365 scenes with pixel-level object annotations. The process involves analyzing the activations of the dense layer neurons by identifying positive and negative example sets—images that strongly activate a particular neuron and those that do not. Using the object annotations from the dataset, which are mapped to classes in the Wikipedia-derived background knowledge base, we employed Concept Induction to generate descriptive labels corresponding to human-understandable concepts for each neuron.

We conducted extensive evaluations, including statistical analyses of neuron activations and Concept Activation Analysis (led by Rushrukh Ryan), to assess the relevance of the extracted concepts across the hidden layer activation space. Our method was compared against state-of-the-art Explainable AI techniques, namely CLIP-Dissect (concept extraction led by Rushrukh Ryan) and GPT-4 (concept extraction led by Adrita Barua), which utilize predefined concept pools or large language models for explanation generation. The results demonstrate that our method not only provides more meaningful and diverse explanations but also achieves competitive or superior performance in both quantitative and qualitative assessments. Specifically, our approach yields a higher number of high-relevance concepts and shows statistically significant associations between neuron activations and the generated concepts. Although our experiments focused on CNNs and image scene classification using the ADE20K dataset, the method is inherently model-agnostic and can be applied to various neural network architectures and data modalities. The customizable background knowledge base allows the approach to be adapted to specialized domains, such as medical diagnosis or

other fields with domain-specific knowledge graphs.

Despite its strengths, the approach relies on the availability of extensive labeled data and a rich background knowledge base, which may limit its applicability in domains with limited resources. Additionally, the method focuses on dense layer activations, potentially overlooking nuanced feature interactions in earlier layers of the neural network. However, these limitations present opportunities for future research, such as integrating domain-specific knowledge bases and extending the analysis to multiple layers within the network.

[35] Samatha Ereshi Akkamahadevi, Abhilekha Dalal, and Pascal Hitzler. *Automating CNN Neuron Interpretation using Concept Induction.* In International Semantic Web Conference, 2024. To appear.

This publication *Automating CNN Neuron Interpretation using Concept Induction,*[35] co-authored with Samatha Ereshi Akkamahadevi, presents a fully automated pipeline designed to assign human-understandable interpretations to the activations of hidden neurons in deep learning models. Building on previous work that manually employed Concept Induction and semantic reasoning with a Wikipedia-derived concept hierarchy, this study enhances the process by automating it, significantly reducing execution time from several days to approximately one hour and fifteen minutes while ensuring consistent results.

The key contribution of this work lies in the automation of essential stages, including model training, Concept Induction, image retrieval, and statistical validation, thereby eliminating the need for manual intervention. Utilizing Concept Induction generates interpretable class labels for hidden neurons, enabling efficient and understandable interpretations of neural activations. This not only enhances transparency but also demonstrates robustness through rigorous statistical analysis. Furthermore, automation ensures reproducibility, thereby advancing the scalability of explainable deep learning without compromising interpretability.

Abhilekha Dalal assisted Samatha Ereshi Akkamahadevi in developing and optimizing the automated pipeline, contributing to the implementation of parallel processing techniques and the integration of various system stages. This collaboration ensured the robustness and efficiency of the approach, demonstrating the potential of automated, scalable methods to

enhance the interpretability of deep learning models, ensuring consistent and reproducible results, thereby facilitating broader application and trust in deep learning systems.

Chapter 5

Error-margin Analysis for Hidden Neuron Activation Labels

5.1 Overview

Building upon our work in Chapter 4, where we developed an approach to assign human-understandable concepts to hidden neuron activations in deep learning models, we now turn our attention to a critical aspect of this interpretation: the precision of these concept assignments. While our previous efforts significantly enhanced model transparency by providing meaningful labels for neuron activations, we recognized that this was only the first step in truly understanding how high-level concepts are represented within artificial neural networks.

In the field of explainable AI (XAI), the focus has predominantly been on identifying what stimuli activate a neuron in most cases, which corresponds to the notion of recall in information retrieval. However, this approach, while valuable, tells only part of the story. It became evident that to fully grasp the nuances of neural network decision-making, we needed to investigate not just what activates a neuron, but also how often it activates incorrectly – in other words, its precision.

This realization led us to consider the concept of "error margins" for neuron labels. By understanding how neurons respond to stimuli that do not correspond to their assigned

labels (non-target labels), we can potentially quantify the likelihood of false positives. This additional layer of information is crucial for several reasons:

1. Enhanced Interpretability: Understanding both the recall and precision of neuron-concept associations could provide a more complete picture of how information is represented within the network.

2. Practical Usefulness: For neuron labeling to be practically applicable in real-world scenarios, users need to know not just what a neuron activation might indicate, but also how reliable that indication is.

3. Distributed Representation: Neural networks naturally employ distributed representations, where the simultaneous activation of multiple neurons conveys information. Error margin analysis could help us understand this complex interplay more accurately.

4. Trust and Accountability: In critical applications where explanations are crucial for trust, knowing the error margins of concept detections could add an essential layer of accountability.

The motivation behind this work stems from the need to move beyond simple labeling of neurons to a more nuanced understanding of their behavior. By investigating the precision of neuron-concept associations, we aim to provide a more robust and reliable framework for interpreting neural network decisions. This approach has the potential to not only enhance our understanding of how neural networks process information but also to improve the trustworthiness and applicability of AI systems in critical domains. Hence, we formulate the following research question: *How can we quantify and validate the error margins of concept-based neuron labels in deep neural networks to enhance the precision and reliability of explainable AI techniques?*

5.2 Contributions

[36] Abhilekha Dalal, Rushrukh Rayan, and Pascal Hitzler. *Error-margin analysis for hidden neuron activation labels.* In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, Neural-Symbolic Learning and Reasoning, pages 149–164, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.

In this publication *Error-margin analysis for hidden neuron activation labels.*[36], we address the research question of enhancing the interpretability of deep learning models by introducing a novel framework for error-margin analysis of neuron activations. Building on prior work involving concept-based neuron labeling, this chapter provides a structured way to quantify the reliability of neuron-concept associations, ensuring that explanations reflect not only accurate activations but also avoid false positives.

Our method introduces the concept of Non-Target Label Activation (Non-TLA) percentages to evaluate how frequently neurons incorrectly activate for non-target concepts. This error-margin framework ensures that explanations are not only insightful but also precise, offering end-users greater confidence in AI outputs. We develop a comprehensive evaluation framework that leverages both Google Images and the ADE20K dataset to validate our error margin calculations. This dual-dataset approach ensures the robustness and generalizability of our findings across different image sources.

Our work employs statistical methods, including Mann-Whitney U tests and Wilcoxon signed-rank tests, to validate the significance of our error margin calculations. This statistical rigor enhances the reliability and trustworthiness of our results. We extend our analysis beyond individual neurons to examine neuron ensembles, providing insights into how distributed representations in neural networks contribute to concept detection and interpretation. The results of our analysis revealed significant insights into the behavior of neural networks. We found that the error margins derived from the Google Images dataset generalized well to the more structured ADE20K dataset, with consistently lower non-target label activations (Non-TLA) observed in the latter. This trend was statistically significant

across various activation thresholds, confirming the reliability of our error margin calculations. Notably, we identified several concepts with high precision (low error margins) across datasets, providing valuable information about which neuron-concept associations are most trustworthy.

By introducing error margins as a new metric in explainable AI, we provide a more nuanced and complete picture of neural network decision-making processes, going beyond simple concept labeling. Our findings have direct implications for improving the trustworthiness and applicability of AI systems in critical domains, offering a quantitative measure of confidence in concept detection tasks.

[37] Abhilekha Dalal and Pascal Hitzler. Conceptlens: from pixels to understanding, 2024. URL https://arxiv.org/abs/2410.05311.

This publication *ConceptLens: From Pixels to Understanding*[37], makes a critical contribution to addressing the research question stated earlier in the section 5.1 by introducing a visualization tool that simplifies complex neural interpretations into user-friendly, easily accessible insights.

A key motivation behind ConceptLens is the need to bridge the gap between the complexity of deep neural networks (DNNs) and the practical usability of their explanations, particularly for non-expert users. Traditional approaches in explainable AI (XAI) often generate technical results that require substantial domain expertise to interpret. However, in real-world applications—ranging from healthcare to finance—stakeholders need to trust and understand how AI models arrive at decisions. A tool that translates intricate neuron behavior into intuitive visualizations, accessible to users with varying levels of expertise, is essential for building trust and transparency in AI systems. ConceptLens addresses this challenge by converting complex internal processes into visual outputs—bar charts that display neuron activations and their associated error margins—making the system's behavior interpretable at a glance.

At its core, ConceptLens integrates deep learning with symbolic reasoning techniques, employing Concept Induction to assign meaningful labels to hidden neurons based on a vast Wikipedia-derived knowledge base. However, the underlying complexity of this method goes

far beyond simple neuron-label associations. It requires analyzing hidden neuron activations, mapping them to high-level semantic concepts, and calculating error margins to quantify the likelihood that a neuron activation corresponds correctly to the assigned label. This error-margin analysis ensures that the system accounts for both recall (what stimuli activate neurons) and precision (how reliably neurons respond to relevant stimuli), addressing the inherent problem of false positives. The tool validates that the error margins are reliable across different contexts by applying these techniques to multiple datasets, such as ADE20K and Google Images.

While this approach offers powerful insights into neural behavior, its complexity could render the explanations inaccessible to end-users. ConceptLens overcomes this barrier by packaging these technical results into real-time visual feedback through a user-friendly interface. Users can upload images and instantly see bar chart visualizations of concept activations along with their associated error margins, providing both clarity and transparency. This ability to visualize complex results in an intuitive way ensures that not only experts but also non-specialists can confidently interpret the network's behavior.

In summary, ConceptLens addresses the need for both interpretability and accessibility by simplifying complex neural explanations into intuitive visual outputs. Its integration of deep learning and symbolic reasoning techniques ensures that concept-based explanations are precise, while error-margin analysis adds an essential layer of reliability. This contribution aligns closely with the research question by offering a practical, scalable, and user-friendly tool that enhances both the transparency and trustworthiness of AI systems.

Chapter 6

Conclusion

6.1 Summary

In the domain of Explainable AI, a persistent challenge lies in the accurate interpretation of hidden neuron activations. Precise interpretations hold the potential to unveil the inner workings of deep learning systems, shedding light on what these systems internally recognize as relevant within their inputs. This clarity helps us to understand deep learning models, which are often difficult to interpret. While prior research has hinted at the interpretability of hidden neuron activations through different methods, the development of systematic, automated methods for generating and validating these interpretations is largely uncharted territory.

This thesis presented a comprehensive framework for enhancing the interpretability of deep neural networks using structured background knowledge and deductive reasoning. Throughout this dissertation, we have outlined several contributions that have advanced and refined the process to its current state. First, we developed a novel model-agnostic post-hoc Explainable AI (XAI) framework that utilized structured background knowledge to interpret the hidden neurons from dense layer of CNN architecture. By utilizing the Concept Induction system over a Wikipedia-derived concept hierarchy, we automatically generated human-understandable labels for hidden neuron activations, bridging the gap be-

tween deep neural network outputs and human comprehension. This approach outperformed existing XAI approaches in both quantitative and qualitative evaluations, providing a more intuitive understanding of what neurons represent. Second, we addressed the issue of how reliable these neuron labels are by introducing error-margin analysis. Traditional approaches often overlook the reliability of neuron activations, which can lead to misleading or ambiguous explanations. Our error-margin analysis provided a statistical assessment of how often neurons activated incorrectly for non-target labels, significantly reducing false positives and enhancing the trustworthiness of concept-based interpretations. This combination of concept labeling and error-margin analysis ensured for a more detailed and accurate understanding of deep learning models.

Another significant contribution is the development of a fully automated end-to-end system for interpreting neurons' activations using Concept Induction, significantly reducing execution time and ensuring reproducibility. Additionally, the creation of ConceptLens, a visualization tool that translates complex neural interpretations into user-friendly, easily accessible insights, bridges the gap between technical explanations and practical usability for non-expert users.

These contributions collectively address the challenges of the "black box" nature of deep learning models by providing interpretable, reliable, and accessible explanations of neural network decision-making processes. The integration of symbolic reasoning with deep learning techniques offers a promising direction for enhancing the transparency and trustworthiness of AI systems, particularly in critical domains where explainability is crucial.

6.2 Future Work

While the proposed framework has made significant strides in improving the interpretability of deep learning models, there are several avenues that can be explored further to advance this promising direction. For example, I would like to propose looking into the following -

1. *Expanding to Other Neural Architectures*: Our approach focused primarily on Convolutional Neural Networks (CNNs) for image scene classification. Future work could ex-

tend this framework to other neural architectures, such as Recurrent Neural Networks (RNNs) and Transformers, to explore how the interpretability techniques translate across different domains and modalities.

2. *Domain-Specific Knowledge Graphs*: The current implementation utilizes a Wikipedia-derived concept hierarchy as background knowledge. Future research could investigate the integration of domain-specific knowledge graphs, such as medical or financial ontologies, to enhance the applicability of the approach in specialized fields. This would allow for more precise and contextually relevant explanations in domains that require domain-specific expertise.

3. *Layer-wise Interpretability*: Our work primarily focused on interpreting hidden neuron activations in the dense layers of CNNs. Future research could explore interpretability across all layers, including convolutional and pooling layers, to provide a holistic understanding of how information is processed throughout the network.

4. *Integration with active learning:* Exploring ways to incorporate the generated explanations into active learning frameworks could potentially improve model training efficiency and performance while maintaining interpretability.

5. *Addressing adversarial vulnerabilities:* Investigating how concept-based explanations and error margin analysis can be leveraged to enhance model robustness against adversarial attacks represents an important direction for future research

By pursuing these research directions, future work can build upon the foundations laid in this thesis to further advance the field of explainable AI, ultimately leading to more transparent, trustworthy, and widely applicable artificial intelligence systems.

Chapter 7

Parallel Research Initiatives and Collaborative Side Projects

In this chapter, we will describe research initiatives that are being built upon the work outlined in this proposal, expanding through parallel efforts and will also highlight collaborative side projects that we have been part of. This chapter serves to provide essential context for understanding the breadth and impact of the research work within the proposal thesis.

7.1 Parallel Efforts

7.1.1 Automating CNN Neuron Interpretation using Concept Induction

Primary Author:- Samatha Ereshi Akkamahadevi (Master's Thesis)

In this work, we contribute a fully automated pipeline for interpreting hidden neuron activations in Convolutional Neural Networks (CNNs), which represents a key challenge in Explainable AI (XAI). Our previous manual approach, which utilized Concept Induction and semantic reasoning through a Wikipedia-derived concept hierarchy, required several days to complete. With this new automation pipeline, we have significantly streamlined the work-

flow, enabling consistent results while drastically reducing execution time. Specifically, our contributions include the automation of model training, data preparation, Concept Induction, image retrieval, classification, and statistical validation, completely eliminating manual intervention. The developed pipeline not only enhances efficiency but also provides a scalable mechanism to modify parameters—such as expanding training images and classes—and facilitates deeper exploration of Concept Induction results across multiple neuron layers through diverse analytical tools

7.1.2 Concept Induction Using Large Language Models

Primary Author:- Adrita Barua

In the broader context of cognitive processes, Concept Induction stands as a fundamental ability that enables the discernment of meaningful distinctions among objects, categories, and themes. This work builds upon the foundation laid out in the proposal by focusing on the potential of advanced Language Models (LMs), particularly exemplified by ChatGPT, as a Concept Induction system for image sets, with the aim of replacing the Concept Induction module in neural networks.

To assess ChatGPT's performance in Concept Induction, we compare its explanations for distinctions among image groups to those generated by the ECII model, a logical Concept Induction system, and by humans. Participants in a survey task evaluate and compare these explanations, aiming to determine the alignment between ChatGPT's responses and human judgments.

The primary aim of this work is to assess ChatGPT's alignment with human judgments and its potential to outperform other Concept Induction systems like ECII. The work delves into ChatGPT's strengths and limitations in Concept Induction and explores the role of prompt engineering in enhancing its performance. Through this exploration, the aim is to gain a deeper understanding of how language models, such as ChatGPT, navigate the intricate Concept Induction landscape, ultimately advancing AI's capacity for comprehending and articulating complex relationships to improve explainability.

7.1.3 Explainability of Deep Learning Models

Primary Author:- Rushrukh Ryan

This research focuses on automating the extraction of meaningful concepts while expanding the scope to encompass complex and abstract domains. The proposed work introduces a Concept Classifier workflow trained on hidden layer activations from positive and negative samples of the concept. Once trained, the concept classifier can be used to predict the presence or absence of the concept in a new data point.

The proposed solution leverages Concept Induction Analysis and Concept Activation Analysis. Validation metrics, including Concept Accuracy, Statistical Significance, Concept Size Impact, Global Explanations, and Kernel Sensitivity, are employed to evaluate the approach.

By automating concept extraction and relevance assessment within neural network activations, this work provides a systematic framework for advancing automated knowledge discovery and concept understanding.

7.1.4 Interpretability for DeepFake Image Dataset

Primary Author:- Moumita and Samatha Ereshi Akkamahadevi

In this extended phase of our research, we aim to apply the method outlined in this proposal to AI-generated deepfake image datasets to generate explanations for these synthetic images and compare the results with those from the original ADE20K dataset. The research seeks to understand how well our interpretability model performs on artificial versus real images, potentially uncovering insights into the model's ability to detect and explain deepfakes. This work could have significant implications for improving AI interpretability in the context of synthetic media and deepfake detection.

7.2 Collaborative Side Projects

7.2.1 KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies

This project aims to enhance data-driven decision-making, especially in the realm of geographic data analytics, by creating the KnowWhereGraph - a groundbreaking knowledge graph tool. It introduces innovations such as an open, web-accessible knowledge graph for collaborative contributions, the integration of imagery data, spatially-informed AI methods, and user-friendly interfaces for enhanced geographic data analysis.

Contributions:- included adapting and connecting various ontologies to the KWG through triplification process. Additionally, there was a pivotal role in maintaining and improving two major datasets - Wildfire and Cropland.

Publications:-

1. Cogan Shimizu, Shirly Stephen, Rui Zhu, Kitty Currier, Mark Schildhauer, Dean Rehberger, Krzysztof Janowicz, Pascal Hitzler, Colby Fisher, Mohammad Saeid Mahdavinejad, Antrea Christou, Adrita Barua, Abhilehka Dalal, Sanaz Saki Norouzi, Yuanyuan Tian, Ling Cai, Gengchen Mai, Zhangyu Wang, Zilong Liu and Meilin Shi, The KnowWhereGraph Ontology: A Showcase. 13th International Conference on Formal Ontology in Information Systems (FOIS) 2023.

2. Cogan Shimizu, Shirly Stephen, Antrea Christou, Kitty Currier, Mohammad Saeid Mahdavinejad, Sanaz Saki Norouzi, Abhilekha Dalal, Adrita Barua, Colby Fisher, Anthony D'Onofio, Thomas Thelen, Krzysztof Janowicz, Dean Rehberger, Mark Schildhauer and Pascal Hitzler, KnowWhereGraph-Lite: A Perspective of the KnowWhereGraph. In the Knowledge Graph and Semantic Web Conference (KGSWC) 2023.

7.2.2 BioWRAP (Bioplastics With Regenerative Agricultural Properties): Spray-on bioplastics with growth synchronous decomposition and water, nutrient, and agrochemical management

The objective is to develop customizable, locally-sourced biopolymer-based films that can be sprayed onto soil surfaces. These films will be tailored to match crop growth cycles in various climatic conditions and applied with precision agricultural equipment. The spray-on biopolymer mulch will serve as a weed-suppressing barrier, reducing the need for herbicides and mitigating selection pressure on herbicide-resistant weed populations.

Contribution:- literature review for the paper

Publication:-

1. Michael Madin, Katherine Nelson, Kanij Fatema, Karina Schoengold, Abhilekha Dalal, Ikenna Onyekwelu, Rushrukh Rayan, Sanaz Saki Norouzi, Synthesis of current evidence on factors influencing the suitability of synthetic biodegradable mulches for agricultural applications: A systematic review, Journal of Agriculture and Food Research, Volume 16, 2024, 101095, ISSN 2666-1543, https://doi.org/10.1016/j.jafr.2024.101095.

7.2.3 GRIPX: Towards a Global Food systems Data Hub: Seeding the Center for Sustainable Wheat Production

In this project, we collaborate with stakeholders to create a Knowledge Graph (KG) that emphasizes sustainable intensification of wheat production. This marks the initial phase in establishing a Global Food Systems Data Hub and launching a Center for Sustainable Wheat Production. We will demonstrate the KG's utility by showcasing three wheat-related use cases pertinent to the Kansas agricultural sector and the development of a publicly accessible Data Hub.

Contribution:- - The project is in its initial phase, where some preliminary steps are

necessary before we can proceed to create a knowledge graph from the collected data and apply deep learning techniques to it.

Bibliography

[1] Muthukrishnan Ramprasath, M Vijay Anand, and Shanmugasundaram Hariharan. Image classification using convolutional neural networks. *International Journal of Pure and Applied Mathematics*, 119(17):1307–1319, 2018.

[2] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. The Proceedings of Machine Learning Research, 2014.

[3] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1044–1054, 2013.

[4] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550: 354–359, 2017.

[5] Hyuk-Il Choi, Seok-Ki Jung, Seung-Hak Baek, Won Hee Lim, Sug-Joon Ahn, Il-Hyung Yang, and Tae-Woo Kim. Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery. *Journal of Craniofacial Surgery*, 30 (7):1986–1989, 2019.

[6] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[7] Nitin Kanagaraj, David Hicks, Ayush Goyal, Sanju Tiwari, and Ghanapriya Singh. Deep learning using computer vision in self driving cars for lane and traffic sign detection. *International Journal of System Assurance Engineering and Management*, 12(6):1011–1025, 2021.

[8] Isobel Asher Hamilton. Apple cofounder Steve Wozniak says Apple Card offered his wife a lower credit limit. *Business Insider*, November 2019. URL https://www.businessinsider.com/apple-card-sexism-steve-wozniak-2019-11. https://www.businessinsider.com/apple-card-sexism-steve-wozniak-2019-11.

[9] Jake Silberg and James Manyika. Tackling bias in artificial intelligence (and in humans). *McKinsey Global Institute*, June 2019. URL https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.

[10] Sanchari Dhar and Lior Shamir. Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. *Visual Informatics*, 5(3):92–101, 2021. ISSN 2468-502X. doi: https://doi.org/10.1016/j.visinf.2021.10.001.

[11] Zhilu Chen and Xinming Huang. End-to-end learning for lane keeping of self-driving cars. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1856–1860. IEEE, 2017.

[12] Ahmet Sureyya Rifaioglu, Esra Nalbat, Volkan Atalay, Maria Jesus Martin, Rengul Cetin-Atalay, and Tunca Doğan. Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations. *Chemical science*, 11(9):2531–2557, 2020.

[13] Walid Hariri and Ali Narin. Deep neural networks for covid-19 detection and diagnosis using images and acoustic-based techniques: a recent review. *Soft computing*, 25(24):15345–15362, 2021.

[14] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[15] I Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. Manipulating SGD with data ordering attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18021–18032, 2021.

[16] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models: [extended abstract]. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 931–942, 2022. doi: 10.1109/FOCS54457.2022.00092.

[17] Tim Clifford, Ilia Shumailov, Yiren Zhao, Ross Anderson, and Robert Mullins. ImpNet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks, 2022.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[19] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable AI really mean? a new conceptualization of perspectives. In *CEUR Workshop Proceedings*, volume 2071. CEUR, 2018.

[20] Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. Neural-symbolic integration and the semantic web. *Semantic Web*, 11(1):3–11, 2020.

[21] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[23] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[25] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. 2018. URL http://arxiv.org/abs/1806.08049. cite arxiv:1806.08049, presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.

[26] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

[27] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of

Proceedings of Machine Learning Research, pages 3145–3153. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

[28] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frédéric Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/867c06823281e506e8059f5c13a57f75-Abstract-Conference.html.

[29] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*. ICLR, 2023. URL https://openreview.net/forum?id=iPWiwWHc1V.

[30] Jens Lehmann and Pascal Hitzler. Concept learning in description logics using refinement operators. *Mach. Learn.*, 78(1-2):203–250, 2010. doi: 10.1007/s10994-009-5146-2.

[31] Md. Kamruzzaman Sarker, Joshua Schwartz, Pascal Hitzler, Lu Zhou, Srikanth Nadella, Brandon S. Minnery, Ion Juvina, Michael L. Raymer, and William R. Aue. Wikipedia knowledge graph for explainable AI. In Boris Villazón-Terrazas, Fernando Ortiz-Rodríguez, Sanju Mishra Tiwari, and Shishir K. Shandilya, editors, *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*, pages 72–87. Springer, 2020. doi: 10.1007/978-3-030-65384-2_6.

[32] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-

cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[34] Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y. Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. On the value of labeled data and symbolic methods for hidden neuron activation analysis. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, *Neural-Symbolic Learning and Reasoning*, pages 109–131, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.

[35] Samatha Ereshi Akkamahadevi, Abhilekha Dalal, and Pascal Hitzler. Automating cnn neuron interpretation using concept induction. In *International Semantic Web Conference*, 2024. To appear.

[36] Abhilekha Dalal, Rushrukh Rayan, and Pascal Hitzler. Error-margin analysis for hidden neuron activation labels. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, *Neural-Symbolic Learning and Reasoning*, pages 149–164, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-71170-1.

[37] Abhilekha Dalal and Pascal Hitzler. Conceptlens: from pixels to understanding, 2024. URL https://arxiv.org/abs/2410.05311.

[38] Nafiz Shahriar. What is convolutional neural network — cnn (deep learning). https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5, February .

[39] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press, 2010. ISBN 9781420090505. URL http://www.semantic-web-book.org/.

[40] Pascal Hitzler. A review of the semantic web field. *Commun. ACM*, 64(2):76–83, 2021. doi: 10.1145/3397512. URL https://doi.org/10.1145/3397512.

[41] Cara Widmer, Md. Kamruzzaman Sarker, Srikanth Nadella, Joshua Fiechter, Ion Juvina, Brandon S. Minnery, Pascal Hitzler, Joshua Schwartz, and Michael L. Raymer. Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge, 2022. URL https://doi.org/10.48550/arXiv.2209.13710.

[42] Md Kamruzzaman Sarker and Pascal Hitzler. Efficient concept induction for description logics, 2018. URL https://arxiv.org/abs/1812.03243.

[43] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

[44] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.

[45] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[46] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[48] Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Caleb Chen Cao, and Lei Chen. An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM sigkdd conference on knowledge discovery & data mining*, pages 3200–3208, 2021.

[49] Angelos Chatzimparmpas, Rafael M Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020.

[50] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

[51] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020.

[52] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[53] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[55] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[56] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[57] Enric Junque De Fortuny and David Martens. Active learning-based pedagogical rule

extraction. *IEEE transactions on neural networks and learning systems*, 26(11):2664–2677, 2015.

[58] Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. Deepred–rule extraction from deep neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*, pages 457–473. Springer, 2016.

[59] Guido Bologna. A simple convolutional neural network with rule extraction. *Applied Sciences*, 9(12):2411, 2019.

[60] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[61] Vitor Horta and Alessandra Mileo. *Chapter 27. Explaining CNNs Using Knowledge Extraction and Graph Analysis*. 07 2023. ISBN 9781643684062. doi: 10.3233/FAIA230159.

[62] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018.

[63] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[64] Roberto Confalonieri, Tillman Weyde, Tarek R Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

[65] Natalia Díaz-Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco 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*, 79:58–83, 2022.

[66] Md. Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael L. Raymer, and Pascal Hitzler. Explaining trained neural networks with semantic web technologies: First steps. In Tarek R. Besold, Artur S. d'Avila Garcez, and Isaac Noble, editors, *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.

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

[68] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*. ICLR, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

[69] Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, 2022.

Appendix A

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

This publication, which appeared in Neural-Symbolic Learning and Reasoning: 18th International Conference (Nesy 2024), represents the research contribution addressing Objective 1 of this dissertation, with a detailed discussion of contributions and findings presented in Chapter 4.2.

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

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

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

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

Keywords: Explainable AI · Concept Induction · CNN · LLM

1 Introduction

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

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

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

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

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

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

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

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

Core contributions of the paper are as follows.

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

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

2 Concept Extraction

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

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

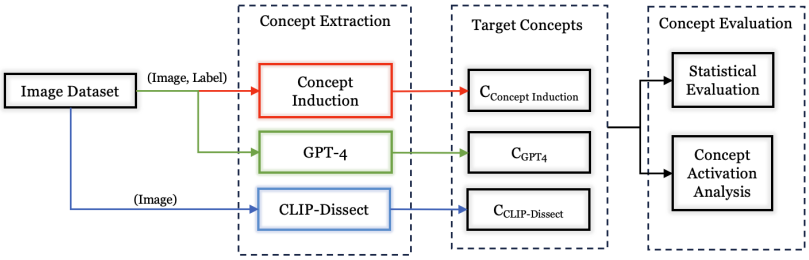

Fig. 1: An overview of the complete pipeline explored in this paper where Concept Extraction outlines the methods used to extract Target Concepts and Concept Evaluation outlines the evaluation methods.

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

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

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

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

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

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

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

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

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

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

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

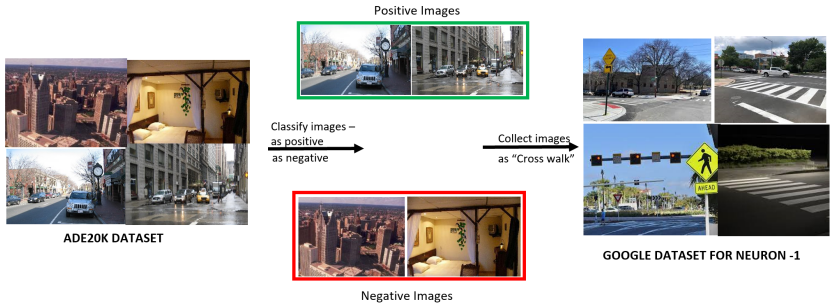

Fig. 2: Example of images that were used for generating and confirming the label hypothesis for neuron 1.

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

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

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

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

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

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

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

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

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

3 Evaluation

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

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

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

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

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

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

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

we consider the label "cross_walk" confirmed for neuron 1 since $88.710 \geq 80$.

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

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

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

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

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

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

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

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

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

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

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

4 Results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5 Discussion

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

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

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

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

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

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

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

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

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

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

6 Conclusion

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

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

References

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

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

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

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

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

33. Sarker, M.K., Xie, N., Doran, D., Raymer, M.L., Hitzler, P.: Explaining trained neural networks with semantic web technologies: First steps. In: Besold, T.R., d'Avila Garcez, A.S., Noble, I. (eds.) Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy). CEUR Workshop Proceedings, vol. 2003. CEUR-WS.org (2017), https://ceur-ws.org/Vol-2003/NeSy17_paper4.pdf
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
35. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? (2016), http://arxiv.org/abs/1611.07450
36. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/shrikumar17a.html
37. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2013), https://api.semanticscholar.org/CorpusID:1450294
38. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p. 180–186. AIES '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375830
39. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. CoRR **abs/1706.03825** (2017), http://arxiv.org/abs/1706.03825
40. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR **abs/1711.00399** (2017), http://arxiv.org/abs/1711.00399
41. Widmer, C., Sarker, M.K., Nadella, S., Fiechter, J., Juvina, I., Minnery, B.S., Hitzler, P., Schwartz, J., Raymer, M.L.: Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge (2022). https://doi.org/10.48550/arXiv.2209.13710
42. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
43. Zarlenga, M.E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept embedding models: Beyond the accuracy-explainability trade-off. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
44. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.P.: Invertible concept-based explanations for cnn models with non-negative concept activation vectors. Proceedings of the AAAI Conference on Artificial Intelligence **35**(13), 11682–11690 (May 2021). https://doi.org/10.1609/aaai.v35i13.17389

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

A Appendices

A.1 Related work

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

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

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

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

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

A.2 Detailed results of Statistical Evaluation

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

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

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

A.3 Detailed results of Concept Activation Analysis

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

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

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

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

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

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

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

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

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

Appendix B

Automating CNN Neuron Interpretation using Concept Induction

This publication, which is to appear at the International Semantic Web Conference (ISWC 2024), represents the research contribution addressing Objective 1 of this dissertation, with a detailed discussion of contributions and findings presented in Chapter 4.2.

Automating CNN Neuron Interpretation using Concept Induction

Samatha Ereshi Akkamahadevi[1,*], Abhilekha Dalal[1] and Pascal Hitzler[1,*]

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

Abstract

This paper presents an automation pipeline for interpreting hidden neuron activations in Convolutional Neural Networks (CNNs), a crucial objective of Explainable AI (XAI). Previously, our research group addressed this objective by employing concept induction and semantic reasoning using a concept hierarchy derived from the Wikipedia knowledge graph. However, the process was executed manually, taking several days to complete. In this study, we have fully automated the workflow, achieving consistent results while significantly reducing the execution time. The automation pipeline streamlines model training, data preparation, concept induction, image retrieval, classification, and statistical validation, thereby completely eliminating the manual intervention. This automation enables us to efficiently interpret and validate CNN neuron activations by modifying parameters, such as incorporating a broader range of training images and classes and examining additional concept induction results across various neuron layers using different analytical tools.

Keywords

Explainable Artificial Intelligence, Deep Learning, Knowledge Graph, Semantic Web, Automation in AI

Demo link: https://youtu.be/a_tHVwexlEE

1. Introduction and Related Work

Deep learning has revolutionized the field of artificial intelligence (AI), achieving breakthroughs in fields such as image recognition, speech recognition, drug discovery, robotics etc. [1]. However, its "black box" nature poses challenges, especially in critical domains needing transparency and explainability [2]. Explainable AI steps in to address these issues, striving to make AI systems more interpretable and their decision-making processes more transparent [3]. Previously, Dalal et al. has demonstrated that hidden neuron activations in CNNs could be meaningfully interpreted using structured background knowledge and ontology reasoning [4, 5, 6]. This approach utilized a large-scale knowledge base derived from Wikipedia's concept hierarchy [7] and employed concept induction [8, 9] to generate interpretable class labels for hidden neurons. Building on this foundation, the current study automates the entire interpretability process to enhance efficiency and ensure reproducibility, eliminating the need for human intervention. We optimized resource allocation and implemented parallel processing to significantly reduce the execution time. This paper provides a detailed description of our automated approach, its technical components, performance evaluation results, and broader implications for XAI.

ISWC 2024 Posters and Demos: 23rd International Semantic Web Conference, November 11–15, 2024, Hanover, MD

*Corresponding author.

samatha94@ksu.edu (S. E. Akkamahadevi); adalal@ksu.edu (A. Dalal); hitzler@ksu.edu (P. Hitzler)

0009-0001-6333-8004 (S. E. Akkamahadevi); 0000-0002-7047-5074 (A. Dalal); 0000-0001-6192-3472 (P. Hitzler)

2. System Architecture and Implementation

2.1. Stage 1: Model Training and Data Configuration

Initially, our automation pipeline trains and configures a CNN model using the ADE20K dataset [10]. This process is executed on Beocat [11], a high-performance computing environment optimized for managing extensive datasets. A Bash script automates job scheduling, resource allocation via SLURM, initializes the Python environment, securely clones the stage 1 repository from GitHub, and installs the necessary dependencies to establish the training environment. We employ a ResNet50V2 architecture implemented in TensorFlow, fine-tuned to enhance model performance using techniques such as data augmentation, early stopping, and batch normalization. Our model is trained on 6,187 images, using Adam optimization algorithm (learning rate 0.001) and categorical cross-entropy as the loss function. Post-training, the model is saved and used to analyze activations within the dense layer across 1,370 ADE20K images and it generates positive and negative example sets based on activation thresholds. P consists of images activating a neuron above 80% of maximum activation, while N includes images activating below 20%. These sets are annotated with classes from background knowledge and will generate configuration files for each neuron which are pivotal for the Concept Induction analysis in Stage 2, providing structured input data for generating and validating label hypotheses.

2.2. Stage 2: Parallelized Concept Induction and Label Hypothesis Generation

We used the concept induction process to generate label hypotheses for each of the 64 neuron activations in the CNN's dense layer using the heuristic Concept Induction system ECII [9]. We automated the simultaneous execution of tasks for all 64 neurons by employing parallel processing with a SLURM-configured Bash script in Beocat. The script initializes the environment, installs necessary Java and Maven dependencies, and clones the latest stage 2 repository from GitHub. Each neuron-specific configuration file from Stage 1 was used to generate semantic concepts, producing output concept files with hypothesized labels and coverage scores using a background knowledge base from the Wikipedia concept hierarchy.

2.3. Stage 3: Parallelized Image Retrieval and Classification

Image retrieval and classification were automated for all neurons to validate the label hypotheses generated in Stage 2. A Bash script manages parallel task execution using SLURM, generating indices for neurons with configuration files. It clones the Stage 3 project repository, sets up the environment, installs dependencies. The script runs a Python program that utilizes the pygoogle_image library to extract labels from the top 3 solutions for each neuron, retrieves 100 images per label from Google, and classifies them using the trained CNN model. Retrieved images are divided into evaluation and verification sets for statistical analysis.

2.4. Stage 4: Statistical Analysis and Verification of Neuron Activations

Label hypotheses are validated through statistical analysis of neuron activations. A Bash script sets up the environment, clones the stage 4 repository, and installs dependencies. The script runs

a Python program that combines activation data from evaluation and verification sets, generates summary statistics, and conducts a Mann-Whitney U test [12] to compare activation values for target and non-target images. Evaluation sets, containing images that strongly activate neurons, provide initial activation metrics. Verification sets undergo further statistical testing to confirm the accuracy and robustness of the label hypotheses.

3. Results and Conclusion

The automation pipeline, executed to enhance the interpretation of hidden neuron activations in CNNs, achieved significant performance improvements.

In stage 1, it eliminated the need for manual analysis to identify and categorize the positive and negative images from the model output. It also generated neuron-specific configuration files with embedded ontology references in OWL format, which serve as input for the subsequent concept induction analysis, completing the execution under 40 minutes.

In stage 2, parallel execution of ECII tool for all 64 neurons reduced the concept induction execution time from over 10 hours to 20 minutes. The ECII tool processed neuron-specific configuration files to generate output concept files with hypothesized labels, sorted by coverage scores, along with precision, recall, and f-measure metrics.

In Stage 3, the image retrieval and classification processes were automated to run concurrently for all neurons to validate label hypotheses from Stage 2. It extracted labels from ECII output, retrieved relevant images from the internet, and classified them using our trained CNN model from stage 1. Model generated the evaluation sets to include activations from images that activate the neuron, providing initial insights into neuron activation patterns while verification sets were generated for detailed statistical analysis in the next stage. This parallelized approach reduced processing time to about 10 minutes, compared to 16 hours without parallelization.

Finally, Stage 4 performed statistical analysis and validated the results in just 3 minutes. The system analyzed activation data, generated a detailed summary of statistics, and verified the label hypotheses. The statistical analysis showed that concept induction analysis with structured background knowledge yields meaningful labels that consistently explain neuron activation. The Mann-Whitney U test rejected the null hypothesis ($p < 0.05$), confirming significant differences in activation values between target and non-target images.

Overall, the entire pipeline was completed in approximately 1 hour 15 minutes, demonstrating substantial improvements in performance, indeed the explainability of the CNN model. The automation and parallelization strategies drastically reduced execution times, minimized manual effort, and ensured consistent and reproducible results, demonstrating the robustness and efficiency of our approach.

4. Future work

We will expand and diversify the dataset, explore various neural network architectures, and integrate various analytical tools. Additionally, we aim to enhance model interpretability by examining additional concept induction results across various neuron layers.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[2] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[3] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xai—explainable artificial intelligence, Science robotics 4 (2019) eaay7120. doi:10.1126/scirobotics.aay7120.

[4] A. Dalal, M. K. Sarker, A. Barua, E. Vasserman, P. Hitzler, Understanding CNN hidden neuron activations using structured background knowledge and deductive reasoning, arXiv preprint arXiv:2308.03999 (2023). doi:10.48550/arXiv.2308.03999.

[5] 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, arXiv preprint arXiv:2404.13567 (2024). doi:10.48550/arXiv.2404.13567.

[6] A. Dalal, R. Rayan, P. Hitzler, Error-margin analysis for hidden neuron activation labels, arXiv preprint arXiv:2405.09580 (2024). doi:10.48550/arXiv.2405.09580.

[7] B. Villazón-Terrazas, F. Ortiz-Rodríguez, S. M. Tiwari, S. K. Shandilya (Eds.), Wikipedia knowledge graph for explainable AI, 2020. doi:10.1007/978-3-030-65384-2_6.

[8] 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.

[9] Efficient concept induction for description logics, volume 33, 2019. doi:10.48550/arXiv.1812.03243.

[10] 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.

[11] K. Hutson, D. Andresen, A. Tygart, D. Turner, Managing a heterogeneous cluster, in: Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning), Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–6. doi:10.1145/3332186.3332251.

[12] P. E. McKnight, J. Najab, Mann-whitney u test, The Corsini encyclopedia of psychology (2010) 1–1. doi:10.1002/9780470479216.corpsy0524.

Appendix C

Error-margin Analysis for Hidden Neuron Activation Labels

This publication, which appeared in Neural-Symbolic Learning and Reasoning: 18th International Conference (Nesy 2024), represents the research contribution addressing Objective 2 of this dissertation, with a detailed discussion of contributions and findings presented in Chapter 5.2.

Error-margin Analysis for Hidden Neuron Activation Labels

Abhilekha Dalal[1], Rushrukh Rayan[1], and Pascal Hitzler[1]

Kansas State University, Manhattan, KS, USA

Abstract. Understanding how high-level concepts are represented within artificial neural networks is a fundamental challenge in the field of artificial intelligence. While existing literature in explainable AI emphasizes the importance of labeling neurons with concepts to understand their functioning, they mostly focus on identifying what stimulus activates a neuron in most cases; this corresponds to the notion of *recall* in information retrieval. We argue that this is only the first-part of a two-part job; it is imperative to also investigate neuron responses to other stimuli, i.e., their *precision*. We call this the neuron label's *error margin*.

Keywords: Explainable AI · Concept Induction · CNN

1 Introduction

Various explainability AI (XAI) techniques in Deep Learning applications have garnered a lot of traction recently. One of the techniques which has proven its efficacy in explainability is associating high-level human understandable concepts with hidden layer neuron activations [7,19,6,13,10]. It is common in the concept-based XAI literature to hand-pick candidate concepts/labels, like the most frequent 20K English words. Previously, we have shown that Concept Induction can be used to assign meaningful labels to neuron activation from a very large knowledge base of candidate concepts used as background knowledge [7][1] in a scene recognition on images scenario.

Statistical analysis in [7] showed that stimuli (network inputs) corresponding to the labels indeed activate the such-labelled neurons with high probability, i.e., the neuron shows high *recall* (in an information retrieval sense) with respect to its "target label". However, the neuron may also activate on many inputs that do not correspond to the neuron's label (e.g., on other neurons' labels, called "non-target labels"), which in information retrieval terms could be understood as low *precision* of the neuron activation, with respect to the assigned label, or in other words, a high false-positives rate, if neuron labels are taken naively at face value.

This is of course not at all unexpected: It is entirely reasonable to assume that any information conveyed by hidden neuron activations be *distributed*, i.e., neurons naturally react to various stimuli, while specific information is indicated

[1] This is under review at NeSy 2024.

by simultaneous activation of neuron groups. However, in order for neuron labelling as in [7] to be practically useful, one would like to use hidden neuron activations to "read off" what the network has detected wrt. a stimulus, i.e. a high false-positive rate is problematic in this scenario.

Herein, we address this issue as follows. We show that the analysis in [7] makes it possible to assign *error margins* to neuron target labels. If a neuron is activated by a stimulus, then the error margin indicates the likelihood that the stimulus indeed falls under the neuron's target label, and this likelihood can be conveyed to the user. We statistically validate error margins by means of data obtained from a user experiment conducted on Amazon Mechanical Turk.

2 Method

In this section, we outline our technical approach for assessing neuron-label associations through *error-margin* analysis (Non-target Label Activation Percentage, or Non-TLA). Non-TLA represents the percentage of images *not falling under the target label* that activate a neuron that carries the target label as per the prior analysis. Similarly, Target Label Activation Percentage, TLA, represents the percentage of images falling under the target label that activate the neuron that carries the target label.

To obtain error margins, we calculate activation percentages for both target labels and non-target labels per neuron based on Google Images retrieved from the labels as search terms, and we also take into account activation patterns of neuron groups for semantically related labels, analyzing TLA and Non-TLA across different cutoff values. We then use images from the ADE20K dataset [33], with annotations improved thorugh Amazon Mechanical Turk, to statistically validate the error-margins obtained earlier. The experimental setting is the same as [7] which we briefly outline below.

Background Premise The primary objective of [7] is to provide insights into the contributions of hidden layer neurons within a Convolutional Neural Network (CNN) in classification tasks. We explore the task of Scene classification using ResNet50V2 on the ADE20K dataset. After training, we examine the activation patterns of the last hidden layer neurons and assign high-level concepts to each neuron, by making use of OWL-reasoning-based Concept Induction over a background knowledge base derived from Wikipedia, comprising approximately 2 million classes, e.g., neuron 1 gets assigned "cross walk" as label. A statistical analysis in [7] shows that each neuron that gets a label assigned (called the neuron's *target label*) indeed activates particularly on network inputs that fall under this label (e.g., images showing a cross walk, for neuron 1): The neuron as an indicator of the target label concept has high recall. However, as the data in [7] also indicates, neurons also tend to activate for many other inputs, i.e. they tend to have relatively low precision.

Computation of Non-TLA Concept Induction in fact generates a number of concept labels for each neuron unit, ranked by some accuracy measure. Herein, we consider the Top-3 labels (ranked by coverage score) for each of the 64 neurons

Table 1: Selective representation of Non-TLA for Google Image dataset(full version can be found in Appendix 4): The table showcases a refined selection, inclusive of concepts with TLA > 80. Non-t: percentage of non-target label images that activate the neuron(s) associated with the concept being analyzed across various activation thresholds.

Concepts	Neuron	**targ %>0**	Non-target % for different threshold values			
			non-t >0	non-t > 20%	non-t > 40%	non-t > 60%
buffet	62	**83.607**	32.714	12.374	3.708	0.825
building	0	**89.024**	72.328	39.552	12.040	2.276
building	0, 63	**80.164**	43.375	12.314	2.276	0.182
building and dome	0	**90.400**	78.185	45.133	14.643	2.639
central_reservation	43	**95.541**	84.973	57.993	19.734	2.913
tap and shower_screen	36	**86.250**	72.584	32.574	7.836	0.860
teapot and saucepan	30	**81.481**	47.984	18.577	4.367	0.845
wardrobe and air_conditioning	19	**89.091**	65.034	31.795	6.958	1.145
skyscraper	22	**99.359**	54.893	21.914	0.977	0.977
skyscraper	54	**98.718**	70.432	26.851	7.050	0.941
skyscraper	63	**94.393**	51.612	20.618	5.775	1.143
skyscraper	22, 54	**97.165**	47.422	7.910	0.465	0.000
skyscraper	22, 63	**96.947**	36.408	5.521	0.449	0.008
skyscraper	54, 63	**96.074**	37.149	5.594	0.615	0.046
skyscraper	22, 54, 63	**95.420**	29.090	3.023	0.234	0.000
skyscraper	26, 54, 63	**81.134**	16.823	1.975	0.350	0.023
skyscraper	22, 26, 54, 63	**80.589**	13.093	0.872	0.015	0.000

in the dense layer. Using the Target-Label image dataset (each image falls under the target label), the TLA is calculated, and, using a Non-target Label image dataset (none of the images contain the target label), the Non-TLA is calculated, for each neuron at specified activation value thresholds, namely > 0, $> 20\%$, $> 40\%$, and $> 60\%$ of the max activation value that was recorded for each neuron. E.g., (see Table 1), neuron 43 activates at $> 40\%$ of its max activation value in about 19.7% of images *not* showing a central reservation.

Neuron Ensembles for Concept Associations The distribution of input information across simultaneously activated neurons necessitates the investigation of neuron ensemble activations at different cut-off activation values. However, an exhaustive analysis of all neuron ensembles does not scale as even just 64 neurons give rise to 2^{64} possible neuron ensembles. We deal with this by considering only ensembles of neurons that activate for semantically related labels. For example, the concept *building* activates both neurons 0 and 63 (see Table 1); we evaluate all images from Non-target Label image dataset as well as Target Label image dataset *separately*, activating neurons 0 and 63 at the specified cut-off activation values, to calculate TLA and Non-TLA.

In scenarios where a concept activates more than two neurons, our analysis encompasses all possible combinations of pairs, triples, etc., of neurons (see *skyscraper* in Table 1). We then narrow our focus to a list of highly associated concepts corresponding to the neurons (see the Concepts column in Table 1), that demonstrate TLA exceeding 80%, i.e., those neurons with high recall.

Annotations of ADE20K Dataset The analysis just described yields *error-margins* associated with each concept, for each of the chosen activation thresholds listed in Table 1. For example, the concept *buffet* has an *error-margin* of

Table 2: Selective representation of Non-TLA for ADE20K and Google Image dataset (for full version see Appendix 5): Non-t: percentage of non-target label images that activate the neuron(s) associated with the concept being analyzed across various activation thresholds.

Concepts	non-t >0		non-t >20%		non-t >40%		non-t >60%	
	Google	ADE20K	Google	ADE20K	Google	ADE20K	Google	ADE20K
buffet	32.714	40.135	12.374	25.817	3.708	9.470	0.825	1.804
building	43.375	11.458	12.314	5.208	2.276	1.458	0.182	0.000
building and dome	78.185	26.170	45.133	5.893	14.643	0.867	2.639	0.000
central_reservation	84.973	44.893	57.993	34.343	19.734	14.927	2.913	3.816
clamp_lamp and clamp	59.504	27.273	29.229	19.170	9.000	8.300	1.652	1.976
closet and air_conditioning	71.054	30.168	38.491	15.620	10.135	5.513	1.267	1.378
cross_walk	28.241	21.474	6.800	16.391	1.524	9.784	0.521	2.922
edifice and skyscraper	48.761	24.187	21.786	8.453	8.379	1.300	2.229	0.260
faucet and flusher	78.562	56.967	37.862	30.580	12.104	11.097	1.873	1.850

12.374 for the Non-TLA of $> 20\%$: Our analysis suggests the *hypothesis* that at most 12.374% of *non-buffet* images activate the neuron unit 62 at 20% of its max activation value. In other words, the *error-margin* at Non-TLA of $> 20\%$ for the concept *buffet* is 12.374%. If this hypothesis can be substantiated, then upon presentation of a new input to the network, activation of neuron 62 of at least 20% of its max activation value means that a *buffet* has been detected, and that this detection is *wrong* in at most about 12.374% of cases.

In order to substantiate our hypotheses, we analyse neuron activation values for new inputs, more precisely for images taken from the ADE20K dataset that was also used in [7]. We take advantage of the fact that ADE20K images already carry rich object annotations, however we have observed that they are still too incomplete for our purposes. Therefore we made use of Amazon Mechanical Turk via the Cloud Research platform, to add missing annotations from a list of concepts derived from Table 1 to 1050 randomly chosen ADE20K images. Details of the study are in Appendix A.2.

Validating Neuron-Concept Associations

To assess the validity of the *error-margins* retrieved from the Google Image dataset for all concepts in Table 1, we look at activations yielded by ADE20K images, and hypothesize that they are similar or lower (i.e., not higher), for non-target images. Non-TLA are computed across the predefined cut-off activation thresholds. Selected values can be found in Table 2.For example, the central reservation neuron 43 mentioned above activates above its 40% max activation threshold for about 14.9% of ADE20K non-target images (not showing central reservations), while it activates for about 19.7% of Google non-target images.

Both single-neuron and neuron ensemble activations are considered and shown in Table 2.

3 Statistical Evaluation and Results

For a statistical evaluation of our error margin values, we treat each row, representing a concept-error pair at each threshold level, from Table 2, as an individual

Table 3: Selective representation of Statistical Evaluation for *confirmed* concepts - getting p-value <0.05 for MWU (full version can be found at Appendix 6; G: activations for Google Image dataset; A: activations for ADE20K dataset; Non-t: percentage of non-target images activating the associated neuron(s) analyzed across various activation thresholds.

Concepts	G	A	p
	non-t >0		
building	43.4	11.5	<0.05
building and dome	78.2	26.2	<0.05
central_reservation	84.9	44.8	<0.05
Wilcoxon signed rank test (non-t >0)			**<0.05**
	non-t >20 %		
building	12.3	5.2	<0.05
building and dome	45.2	5.8	<0.05
clamp_lamp and clamp	29.3	19.2	<0.05
closet and air_conditioning	38.5	15.6	<0.05
Wilcoxon signed rank test (non-t > 20%)			**<0.05**

Concepts	G	A	p
	non-t >40 %		
building	2.3	1.4	<0.05
building and dome	14.6	0.8	<0.05
central_reservation	19.7	14.9	<0.05
clamp_lamp and clamp	9.1	8.3	<0.05
Wilcoxon signed rank test (non-t > 40%)			**<0.05**
	non-t >60 %		
building	0.1	0.0	<0.05
building and dome	2.6	0.0	<0.05
central_reservation	2.9	3.8	<0.05
Wilcoxon signed rank test (non-t > 60%)			**<0.05**

hypothesis. For example, the *error-margin* (Non-TLA) for the concept "central reservation" under the > 40 threshold constitutes one hypothesis.

This way, we get $33 \times 4 = 132$ hypotheses to test.

We conduct Mann-Whitney U tests (MWU) [16] with the null hypothesis (H0) stating that there is no difference in Non-TLA across both datasets, while the alternative hypothesis (H1) posits that Non-TLA in Google Images is greater than in the ADE20K dataset. We choose the MWU test for its robustness with non-parametric data and its aptitude for comparing distributions of independent samples. As our Non-TLA data may not adhere to normality and we're comparing distinct datasets (Google Images and ADE20K), the MWU test provides a reliable means to analyze differences in Non-TLA.

Table 3 presents a comparison of Non-TLA between the Google Images and ADE20K datasets for all concepts. Each row represents a concept, with columns displaying the percentage of non-target label images activating associated neuron(s) in both datasets. The p-values from the MWU test indicate the significance of differences in Non-TLA between the datasets. The analysis reveals a consistent trend of decreased Non-TLA in the ADE20K dataset compared to Google Images across various threshold categories. Among the 33 hypotheses tested for the category of Non-TLA > 0, 13 were rejected at a significance level of $p < 0.05$. Similarly, for Non-TLA $> 20\%$, 15 hypotheses were rejected at the same significance level. In the case of Non-TLA $> 40\%$, 21 hypotheses were rejected, while for Non-TLA $> 60\%$, 23 hypotheses were rejected, all at a p-value < 0.05. Concepts with p-value < 0.05 are deemed statistically significant and are identified as *confirmed* concepts, subject to further scrutiny for their reliability and potential implications.

After confirming concepts using the MWU, we proceed to validate them further using Wilcoxon signed-rank tests. To calculate the Wilcoxon test, we used an online website calculator available at [2]. We employ the Wilcoxon test, with the

[2] http://www.statskingdom.com/170median_mann_whitney.html

hypothesis that the difference between Non-TLA of ADE20K and Google Image dataset would be less than or equal to zero (H0), while the alternative hypothesis (H1) suggested a decrease in Non-TLA in the ADE20K dataset compared to the Google image dataset. Each threshold serves as an individual hypothesis for the Wilcoxon test, with Non-TLA of the *confirmed* concepts for Google and ADE20K datasets grouped accordingly. For instance, all confirmed Non-TLA > 0 for both datasets constitute one hypothesis, while those $> 20\%$ form another. The p-values, denoting the significance of the test results, are displayed at the bottom of the table. Remarkably, the obtained p-values for each threshold suggest the rejection of the null hypothesis, indicating statistically significant differences in Non-TLA between the datasets when considered separately. A p-value < 0.05 from this test would indicate a statistically significant decrease in Non-TLA in the ADE20K dataset compared to the Google dataset, further strengthening our findings and highlighting that the error estimates from the Google image data hold, or are even bettered by, the ADE20K images.

We also examine all *confirmed* concepts from all thresholds together in the Wilcoxon test with the same alternative hypothesis ((H1) suggested a decrease in Non-TLA in the ADE20K dataset compared to the Google image dataset), which provides a comprehensive overview of the differences in Non-TLA between the Google and ADE20K datasets across various levels of activation thresholds. This approach aggregates the results from individual thresholds, offering a more consolidated perspective on the overall significance of the differences observed. In our analysis, obtaining a p-value of 5.633e-7, which is less than 0.05, implies the rejection of the null hypothesis. This indicates a statistically significant decrease in Non-TLA in the ADE20K dataset compared to the Google Image dataset when considering all thresholds collectively.

4 Discussion

The error margins derived from our analysis significantly enhance the interpretability and reliability of neural networks. These margins provide a quantitative measure of confidence for concept detection in image analysis tasks. For instance, when a neuron associated with a specific concept (e.g., "buffet") activates above a certain threshold, the error margin allows us to estimate the likelihood that the image actually contains that concept.

Our study demonstrates the robustness of error margin methodology across diverse datasets without assuming identical neuron-concept associations between Google Images and ADE20K. Instead, our primary goal was to validate the generalizability of error margins across these distinct datasets. In our experiments, we observed varying neuron-concept associations across datasets. For instance, while neuron 62 prominently associated with 'buffet' in Google Images, its activation pattern in ADE20K showed similarities but also notable differences. These variations stem from differences in dataset characteristics, training specificity, and concept granularity. Importantly, these differences strengthen our findings. The methodology's ability to produce statistically significant results despite

these variations underscores its robustness and broad applicability. This adaptability is crucial for real-world applications requiring reliable concept detection and interpretability across diverse data contexts.

Our statistical analysis, employing Mann-Whitney U and Wilcoxon signed-rank tests, reveals significant differences in non-target label activations (Non-TLA) between the ADE20K dataset and the Google Images dataset. Crucially, the lower Non-TLA values observed in the ADE20K dataset validate our error margins and underscore their reliability. This validation is important for several reasons:

- **Generalizability**: The fact that error margins derived from the Google Images dataset generalize well to the more structured and annotated ADE20K dataset indicates that our method is not confined to a specific dataset. This enhances the broader applicability of our approach.
- **Reliability**: The reduced Non-TLA in the ADE20K dataset suggests that neuron activation patterns are more precise and reliable when tested on a well-annotated dataset. This finding assures that the calculated error margins are robust and can be trusted for practical AI applications.
- **Foundation for Future Work**: Validating our error margins across different datasets provides a strong foundation for future research, encouraging further exploration of neuron activation patterns and their implications for model explainability.

These error margins significantly enhance the interpretability of neural network decisions by quantifying the reliability of neuron-concept associations. This offers a more nuanced understanding of how the network processes information, going beyond simple neuron labeling to provide insights into the degree of certainty with which we can interpret a neuron's activation. Such information is crucial for building trust in AI systems, especially in critical decision-making scenarios.

Furthermore, our error margin analysis can guide the refinement of neural architectures. By identifying neurons or neuron ensembles with high precision for specific concepts, we can inform targeted improvements in network design. For example, architectures could be optimized to enhance the precision of key concept detections, potentially leading to more efficient and accurate models.

In summary, our analysis demonstrates that the concept associations and error margins derived from our method are both reliable and generalizable. These findings contribute significantly to the field of explainable AI by providing a validated approach to understanding and improving the interpretability of neural networks, paving the way for more advanced and trustworthy AI systems.

5 Conclusion

In this study, we addressed the challenge of understanding high-level concepts within neural networks by proposing a methodology to label neurons with concepts, thereby enhancing model interpretability in explainable AI. Our approach

goes beyond identifying activating stimuli for neurons by examining their responses to both target and non-target concepts. Through systematic analysis and empirical validation using datasets like Google Images and ADE20K, we have demonstrated the effectiveness and generalizability of our method. Statistical evaluation confirms the reliability of error margins obtained from Google Images. Notably, we observed consistent trends indicating decreased non-target activations in the ADE20K dataset, highlighting its potential for robust image analysis tasks.

The derived error margins provide a quantitative measure of confidence for concept detection in image analysis tasks. They enhance the interpretability of neural network decisions by offering insights into neuron activation certainty, crucial for trust in AI systems involved in critical decisions. Additionally, our approach can guide the refinement of neural architectures by identifying neurons with high precision for specific concepts, informing targeted improvements for more efficient and accurate models.

Our contributions include reliable labeling of hidden neuron responses, a systematic approach to analyzing neuron responses to target and non-target concepts, and enhancing the transparency and reliability of concept-based explainable AI systems. This work advances the goal of creating more transparent and interpretable deep learning models, essential for critical applications. Future work could integrate error margin analysis with other explainability methods, like saliency maps, for a holistic understanding of neural network decisions. Additionally, studying error margins across different network layers could reveal insights into hierarchical concept learning.

Acknowledgement. The authors acknowledge partial funding under National Science Foundation grant 2119753.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
3. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. CoRR **abs/1806.08049** (2018), http://arxiv.org/abs/1806.08049
4. Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences **117**(48), 30071–30078 (2020)
5. Confalonieri, R., Weyde, T., Besold, T.R., del Prado Martín, F.M.: Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. Artificial Intelligence **296**, 103471 (2021)
6. Crabbé, J., van der Schaar, M.: Concept activation regions: A generalized framework for concept-based explanations. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 2590–2607. Curran Associates, Inc. (2022)

7. Dalal, A., Rayan, R., Barua, A., Vasserman, E.Y., Sarker, M.K., Hitzler, P.: On the value of labeled data and symbolic methods for hidden neuron activation analysis (2024), https://arxiv.org/abs/2404.13567, NeSy 2024, under review
8. Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., Herrera, F.: Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. Information Fusion **79**, 58–83 (2022)
9. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189 – 1232 (2001). https://doi.org/10.1214/aos/1013203451
10. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
11. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095
12. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI – explainable artificial intelligence. Science robotics **4**(37), eaay7120 (2019)
13. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/kim18d.html
14. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)Reliability of Saliency Methods, p. 267–280. Springer-Verlag, Berlin, Heidelberg (2022), https://doi.org/10.1007/978-3-030-28954-6_14
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
16. McKnight, P.E., Najab, J.: Mann-whitney u test. In: The Corsini Encyclopedia of Psychology. Wiley (2010)
17. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review pp. 1–66 (2022)
18. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations. ICLR (2023), https://openreview.net/forum?id=FlCg47MNvBA
19. Oikarinen, T., Weng, T.W.: CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In: International Conference on Learning Representations. ICLR (2023), https://openreview.net/forum?id=iPWiwWHc1V
20. Procko, T., Elvira, T., Ochoa, O., Rio, N.D.: An exploration of explainable machine learning using semantic web technology. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). pp. 143–146. IEEE Computer Society (jan 2022). https://doi.org/10.1109/ICSC52841.2022.00029

21. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
22. Sarker, M.K., Xie, N., Doran, D., Raymer, M.L., Hitzler, P.: Explaining trained neural networks with semantic web technologies: First steps. In: Besold, T.R., d'Avila Garcez, A.S., Noble, I. (eds.) Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy). CEUR Workshop Proceedings, vol. 2003. CEUR-WS.org (2017), https://ceur-ws.org/Vol-2003/NeSy17_paper4.pdf
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
24. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? ArXiv **abs/1611.07450** (2016), https://api.semanticscholar.org/CorpusID:12993310
25. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/shrikumar17a.html
26. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2013), https://api.semanticscholar.org/CorpusID:1450294
27. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p. 180–186. AIES '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375830
28. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. CoRR **abs/1706.03825** (2017), http://arxiv.org/abs/1706.03825
29. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Cybersecurity (2017), https://api.semanticscholar.org/CorpusID:3995299
30. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
31. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.P.: Invertible concept-based explanations for cnn models with non-negative concept activation vectors. Proceedings of the AAAI Conference on Artificial Intelligence **35**(13), 11682–11690 (May 2021). https://doi.org/10.1609/aaai.v35i13.17389
32. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2131–2145 (2018)
33. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)

A Appendices

A.1 Related work

Efforts to demystify deep learning [12,2,17] are ongoing. Methods for explainability can be categorized based on their approach to understanding input data, such as feature summarization [24,21], or the model's internal representation, like node summarization [32,4]. These methods further classify into model-specific [24] or model-agnostic [21] approaches. Some methods rely on human interpretation of explanatory data, such as posing counterfactual questions [29].

Model-agnostic techniques for feature attribution, such as LIME [21] and SHAP [15], aim to elucidate model predictions by assessing the influence of individual features. However, they encounter challenges like explanation instability [3] and susceptibility to biased classifiers [27]. On the other hand, pixel attribution endeavors to comprehend predictions by assigning significance to individual pixels [26,23,28]. Nonetheless, it faces notable limitations, particularly with ReLU activation [25] and adversarial perturbations [14], leading to inconsistencies in interpretability.

Explanations developed by [13,6] employ supervised learning and curated concepts. These methods utilize classifiers on target concepts, with weights representing Concept Activation Vectors (CAVs). Another approach by [10] utilizes image segmentation and clustering for concept selection, albeit potentially losing information and only applicable to visible concepts. [31] proposed enhancements using Non-negative Matrix Factorization to mitigate information loss. Individual Conditional Expectation (ICE) plots [11] and Partial Dependency Plots [9] provide insights into prediction-feature relationships from both local and global perspectives but may struggle with intricate feature interactions.

Previous studies suggest that hidden neurons may represent high-level concepts [32,4], but these methods often require semantic segmentation [30] (resource-intensive) or explicit concept annotations [13].

Some research have utilized Semantic Web data for explaining deep learning models [5,8], and Concept Induction for providing explanations [22,20]. However, their focus was on analyzing input-output behavior, generating explanations for the overall system.

CLIP-Dissect [19] represents work similar to ours, employing a different approach. They utilize the CLIP pre-trained model, employing zero-shot learning to associate images with labels. Label-Free Concept Bottleneck Models [18], building upon CLIP-Dissect, use GPT-4 [1] for concept set generation. However, CLIP-Dissect has limitations that may be challenging to overcome without significant changes to the approach. These include limited accuracy in predicting output labels based on concepts in the last hidden layer and difficulty in transferring to other modalities or domain-specific applications. The Label-Free approach inherits these limitations and may compromise explainability, as it uses a concept derivation method that is not inherently explainable.

A.2 Details of AMT user-study

Using a subset of randomly chosen 1050 ADE20K images, we conducted a user study through Amazon Mechanical Turk using the Cloud Research platform, to annotate images based on a list of concepts derived from Table4.

The study protocol was reviewed and approved by the Institutional Review Board (IRB) at Kansas State University and was deemed exempt under the criteria outlined in the Federal Policy for the Protection of Human Subjects, 45 CFR §104(d), category: Exempt Category 2 Subsection ii. The study was conducted in 35 batches (each batch containing 30 images), with 5 participants per study compensated with $5 for completing the task. The task was estimated to take approximately 40 minutes, equivalent to $7.50 per hour.

For each image, users were presented with a list of concepts (a concise form of concepts from Table 4) to choose from, including buffet, building, building and dome, central_reservation, clamp_lamp and clamp, closet and air-conditioning, cross_walk, edifice and skyscraper, faucet and flusher, field, flusher and soap_dish, footboard and chain, hedgerow and hedge, lid and soap_dispenser, mountain, mountain and bushes, night_table, open_fireplace and coffee_table, pillow, potty and flusher, road, road and automobile, road and car, route, route and car, shower_stall and cistern, Shower_stall and screen_door, skyscraper, slope, tap and crapper, tap and shower_screen, teapot and saucepan, wardrobe and air-conditioning.

Users were allowed to select multiple concepts for each image, indicating all concepts that applied to the given image. These selected concepts were considered annotations for the respective image.

A.3 Detailed result of Non-TLA and Statistical Evaluation

The detailed results of Non-target Label Activation Percentages (Non-TLA) for the Google dataset are meticulously outlined in Table 4. This table presents a carefully curated selection, focusing on concepts and neuron ensembles with Target Label Activation (TLA) exceeding 80%. It offers valuable insights into the percentage of Non-Target Label images activating the neuron(s) associated with the concept under scrutiny across a spectrum of activation thresholds.

Furthermore, Table 5 compares the Non-TLA between the ADE20K and Google Image datasets, highlighting variations in activation across different thresholds and providing a comprehensive view of dataset-specific nuances.

Lastly, the statistical evaluation for *confirmed* concepts, outlined in Table 6, underscores concepts where statistical significance (with a p-value less than 0.05 for the MWU test) has been established. This table offers insights into the percentage of non-target label images activating the associated neuron(s) across diverse activation thresholds, providing valuable information on the robustness of the identified concepts across both the Google and ADE20K datasets.

Table 4: Non-target Label Activation Percentages (Non-TLA) for Google dataset: The table showcases a refined selection, inclusive of concepts and neuron ensembles with targ(et) activation > 80%. Non-t: percentage of non-target images that activate the neuron(s) associated with the concept being analyzed across various activation thresholds.

Concepts	Neuron	**targ %>0**	Non-target % for different threshold values			
			non-t >0	non-t > 20%	non-t > 40%	non-t > 60%
buffet	62	**83.607**	32.714	12.374	3.708	0.825
building	0	**89.024**	72.328	39.552	12.040	2.276
building	0, 63	**80.164**	43.375	12.314	2.276	0.182
building and dome	0	**90.400**	78.185	45.133	14.643	2.639
central_reservation	43	**95.541**	84.973	57.993	19.734	2.913
clamp_lamp and clamp	7	**95.139**	59.504	29.229	9.000	1.652
closet and air_conditioning	19	**86.891**	71.054	38.491	10.135	1.267
cross_walk	1	**88.770**	28.241	6.800	1.524	0.521
edifice and skyscraper	63	**92.135**	48.761	21.786	8.379	2.229
faucet and flusher	29	**95.695**	78.562	37.862	12.104	1.873
field	18	**91.824**	65.333	30.207	8.183	1.656
flusher and soap_dish	56	**90.094**	63.552	29.901	7.695	1.148
footboard and chain	49	**88.889**	66.702	40.399	17.064	4.399
hedgerow and hedge	54	**91.165**	68.527	30.421	7.685	1.352
lid and soap_dispenser	29	**99.237**	78.571	34.989	9.052	1.485
mountain and bushes	16	**87.037**	24.969	10.424	4.666	1.937
mountain and bush	16	**87.037**	24.969	10.424	4.666	1.937
mountain	43	**99.367**	88.516	64.169	23.112	4.326
night_table	3	**90.446**	56.714	27.691	7.691	1.137
open_fireplace and coffee_table	41	**88.525**	16.381	4.325	0.812	0.088
pillow	3	**98.214**	61.250	28.228	7.249	1.001
pillow	50	**99.405**	66.834	24.242	4.101	0.530
pillow	3, 50	**97.605**	46.492	9.634	0.988	0.049
potty and flusher	29	**88.525**	76.830	36.537	10.755	1.932
road and car	51	**98.810**	48.571	25.373	8.399	3.261
road and automobile	51	**92.560**	41.466	16.055	3.301	0.701
road	48	**100.000**	76.789	47.897	18.843	3.803
road	48, 51	**97.099**	44.592	17.727	3.471	0.702
route	48	**100.000**	80.834	51.873	21.034	4.979
route and car	51	**92.628**	47.408	18.871	4.081	1.416
route	48, 51	**94.334**	45.089	18.937	4.809	1.169
shower_stall and cistern	8	**100.000**	53.186	24.788	8.485	1.372
Shower_stall and screen_door	57	**98.496**	31.747	12.876	4.121	1.026
slope	18	**92.143**	64.503	29.976	6.894	1.200
tap and crapper	36	**89.130**	70.606	36.839	13.696	2.511
tap and shower_screen	36	**86.250**	72.584	32.574	7.836	0.860
teapot and saucepan	30	**81.481**	47.984	18.577	4.367	0.845
wardrobe and air_conditioning	19	**89.091**	65.034	31.795	6.958	1.145
skyscraper	22	**99.359**	54.893	21.914	0.977	0.977
skyscraper	54	**98.718**	70.432	26.851	7.050	0.941
skyscraper	63	**94.393**	51.612	20.618	5.775	1.143
skyscraper	22, 26	**82.116**	22.274	3.423	0.292	0.004
skyscraper	26, 54	**82.225**	28.782	5.444	0.703	0.054
skyscraper	22, 54	**97.165**	47.422	7.910	0.465	0.000
skyscraper	22, 63	**96.947**	36.408	5.521	0.449	0.008
skyscraper	26, 63	**81.788**	21.421	3.335	0.534	0.088
skyscraper	54, 63	**96.074**	37.149	5.594	0.615	0.046
skyscraper	22, 26, 54	**81.461**	18.940	2.363	0.169	0.000
skyscraper	22, 26, 63	**81.243**	15.252	1.706	0.184	0.004
skyscraper	22, 54, 63	**95.420**	29.090	3.023	0.234	0.000
skyscraper	26, 54, 63	**81.134**	16.823	1.975	0.350	0.023
skyscraper	22, 26, 54, 63	**80.589**	13.093	0.872	0.015	0.000

Table 5: Non-target Label Activation Percentages (Non-TLA) for ADE20K and Google Image dataset: Non-t: percentage of non-target label images that activate the neuron(s) associated with the concept being analyzed across various activation thresholds.

Concepts	non-t >0		non-t >20%		non-t >40%		non-t >60%	
	google	ADE20K	google	ADE20K	google	ADE20K	google	ADE20K
buffet	32.714	40.135	12.374	25.817	3.708	9.470	0.825	1.804
building	43.375	11.458	12.314	5.208	2.276	1.458	0.182	0.000
building and dome	78.185	26.170	45.133	5.893	14.643	0.867	2.639	0.000
central_reservation	84.973	44.893	57.993	34.343	19.734	14.927	2.913	3.816
clamp_lamp and clamp	59.504	27.273	29.229	19.170	9.000	8.300	1.652	1.976
closet and air_conditioning	71.054	30.168	38.491	15.620	10.135	5.513	1.267	1.378
cross_walk	28.241	21.474	6.800	16.391	1.524	9.784	0.521	2.922
edifice and skyscraper	48.761	24.187	21.786	8.453	8.379	1.300	2.229	0.260
faucet and flusher	78.562	56.967	37.862	30.580	12.104	11.097	1.873	1.850
field	65.333	66.161	30.207	30.043	8.183	10.412	1.656	2.386
flusher and soap_dish	63.552	19.481	29.901	10.035	7.695	3.896	1.148	0.236
footboard and chain	66.702	27.975	40.399	13.671	17.064	5.063	4.399	1.013
hedgerow and hedge	68.527	45.120	30.421	28.390	7.685	13.308	1.352	2.028
lid and soap_dispenser	78.571	57.512	34.989	18.427	9.052	2.817	1.485	0.352
mountain	88.516	45.144	64.169	33.725	23.112	16.115	4.326	3.842
mountain and bushes	24.969	28.331	10.424	16.573	4.666	6.607	1.937	1.904
night_table	56.714	30.534	27.691	15.267	7.691	5.954	1.137	1.679
open_fireplace and coffee_table	16.381	26.139	4.325	10.590	0.812	2.413	0.088	0.268
pillow	46.492	12.500	9.634	3.869	0.988	1.190	0.049	0.149
potty and flusher	76.830	58.410	36.537	24.194	10.755	4.608	1.932	1.152
road	44.592	8.501	17.727	6.955	3.471	4.328	0.702	0.927
road and automobile	41.466	17.604	16.055	14.497	3.301	8.728	0.701	2.811
road and car	48.571	14.815	25.373	11.704	8.399	6.074	3.261	1.333
route	45.089	12.349	18.937	10.241	4.809	5.723	1.169	1.807
route and car	47.408	17.073	18.871	14.204	4.081	7.461	1.416	2.152
shower_stall and cistern	53.186	25.982	24.788	9.700	8.485	4.965	1.372	1.039
Shower_stall and screen_door	31.747	24.910	12.876	14.320	4.121	5.897	1.026	1.203
skyscraper	13.093	3.009	0.872	0.463	0.015	0.231	0.000	0.116
slope	64.503	66.520	29.976	29.967	6.894	9.879	1.200	1.976
tap and crapper	70.606	62.225	36.839	12.861	13.696	4.890	2.511	0.611
tap and shower_screen	72.584	62.621	32.574	13.180	7.836	4.733	0.860	0.607
teapot and saucepan	47.984	23.632	18.577	11.176	4.367	6.519	0.845	1.281
wardrobe and air_conditioning	65.034	30.525	31.795	16.160	6.958	5.525	1.145	0.967

Table 6: Statistical Evaluation for *confirmed* concepts (concepts getting p-value <0.05 for MWU): Non-t: percentage of non-target label images activating the associated neuron(s) analyzed across various activation thresholds.

Concepts	Google	ADE20K	p-values
	non-t >0		
building	43.37468	11.45833	0.018471
building and dome	78.185	26.16984	6.06E-05
central_reservation	84.97336	44.89338	1.75E-66
closet and air_conditioning	71.05416	30.16845	0.009373
edifice and skyscraper	48.76092	24.18726	0.016058
faucet and flusher	78.562	56.96671	9.19E-07
footboard and chain	66.702	27.97468	0.000284
lid and soap_dispenser	78.57143	57.51174	0.00218
pillow	46.49232	12.5	4.21E-23
potty and flusher	76.82974	58.41014	1.39E-07
shower_stall and cistern	53.1865	25.98152	0.016657
tap and crapper	70.60579	62.22494	6.17E-08
tap and shower_screen	72.584	62.62136	0.007024
Wilcoxon signed rank test (non-t >0)			**0.0001221**
	non-t >20 %		
building	12.31365	5.208333	1.72E-17
building and dome	45.13343	5.892548	1.37E-23
clamp_lamp and clamp	29.2287	19.16996	1.57E-07
closet and air_conditioning	38.4913	15.62021	0.000287
edifice and skyscraper	21.78641	8.452536	5.80E-17
faucet and flusher	37.86209	30.57953	1.80E-15
lid and soap_dispenser	34.98939	18.42723	2.74E-15
mountain and bushes	10.42437	16.57335	3.25E-06
pillow	9.634389	3.869048	3.49E-49
potty and flusher	36.53659	24.19355	3.69E-18
Shower_stall and screen_door	12.87584	14.3201	0.035051
skyscraper	0.872071	0.462963	1.99E-05
tap and crapper	36.83933	12.86064	0.000114
tap and shower_screen	32.5745	13.17961	3.22E-14
wardrobe and air_conditioning	31.79496	16.16022	2.18E-11
Wilcoxon signed rank test (non-t > 20%)			**0.0004272**
	non-t >40 %		
building	2.27609	1.458333	3.16E-19
building and dome	14.64338	0.866551	6.28E-20
central_reservation	19.73357	14.92705	1.18E-05
clamp_lamp and clamp	9.000096	8.300395	2.79E-31
closet and air_conditioning	10.1354	5.513017	6.38E-09
cross_walk	1.52392	9.78399	0.000572
edifice and skyscraper	8.37939	1.30039	5.06E-17
faucet and flusher	12.10377	11.09741	2.90E-24
field	8.183384	10.41215	3.82E-05
flusher and soap_dish	7.695067	3.896104	4.26E-08
lid and soap_dispenser	9.052334	2.816901	2.04E-19
mountain and bushes	4.666314	6.606943	1.28E-12
pillow	0.988239	1.190476	1.37E-23
potty and flusher	10.75519	4.608295	1.97E-09
road	3.471037	4.327666	0.033105
road and car	8.399088	6.074074	0.009958
Shower_stall and screen_door	4.120976	5.89651	1.13E-07
skyscraper	0.015367	0.231481	2.47E-30
slope	6.893903	9.879254	1.14E-07
tap and shower_screen	7.835857	4.73301	2.05E-12
wardrobe and air_conditioning	6.9579	5.524862	1.70E-19
Wilcoxon signed rank test (non-t > 40%)			**0.0479**

	non-t > 60%		
building	0.182087	0	1.08E-07
building and dome	2.639495	0	5.70E-10
central_reservation	2.912966	3.815937	1.50E-07
clamp_lamp and clamp	1.652099	1.976285	4.24E-19
closet and air_conditioning	1.266925	1.378254	2.50E-07
cross_walk	0.520833	2.92249	0.000171
edifice and skyscraper	2.228561	0.260078	4.80E-07
faucet and flusher	1.872623	1.849568	0.008524
field	1.655819	2.386117	1.43E-09
flusher and soap_dish	1.147982	0.236128	3.03E-13
lid and soap_dispenser	1.485149	0.352113	3.10E-07
mountain and bushes	1.936961	1.903695	9.96E-12
pillow	0.048848	0.14881	1.04E-09
potty and flusher	1.931664	1.152074	0.010232
road	0.701794	0.927357	0.000445
road and car	3.261441	1.333333	3.79E-05
route and car	1.415601	2.15208	0.000137
shower_stall and cistern	1.372089	1.039261	0.031085
Shower_stall and screen_door	1.025822	1.203369	9.36E-11
skyscraper	0	0.115741	6.15E-26
slope	1.200192	1.975851	2.39E-10
tap and shower_screen	0.859795	0.606796	3.67E-08
wardrobe and air_conditioning	1.144971	0.966851	1.52E-14
Wilcoxon signed rank test (non-t > 60%)			**0.05803**

Appendix D

ConceptLens: from Pixels to Understanding

This publication represents the research contribution addressing Objective 2 of this dissertation. For a detailed discussion of contributions and findings, please refer to Chapter 5.2.

ConceptLens: from Pixels to Understanding

Abhilekha Dalal[1] and Pascal Hitzler[1]

Kansas State University, Manhattan, KS, USA

Abstract. ConceptLens is an innovative tool designed to illuminate the intricate workings of deep neural networks (DNNs) by visualizing hidden neuron activations. By integrating deep learning with symbolic methods, ConceptLens offers users a unique way to understand what triggers neuron activations and how they respond to various stimuli. The tool uses error-margin analysis to provide insights into the confidence levels of neuron activations, thereby enhancing the interpretability of DNNs. This paper presents an overview of ConceptLens, its implementation, and its application in real-time visualization of neuron activations and error margins through bar charts.

Keywords: Explainable AI · Concept Induction · CNN

1 Introduction

Deep neural networks (DNNs) have revolutionized various fields by offering powerful solutions for complex tasks such as image recognition, natural language processing, and more [8,5,1]. However, the "black box" nature of these models often leaves users and researchers in the dark about how specific decisions are made [6,11]. Understanding the internal workings of these networks is crucial for improving their reliability and trustworthiness. A promising approach to make them more interpretable is to associate the activations of neurons in their hidden layers with human-understandable concepts [2,7,4]. Prior work [2] has focused on identifying the concepts that maximally activate each neuron – corresponding to the notion of recall. However, solely optimizing for recall is insufficient, as neurons tend to also activate for many other inputs that do not match their assigned concept (low precision).

To address this limitation, we present a visualizing tool, *ConceptLens*, that quantifies the uncertainty and imprecision in neural concept labels through error margins. *ConceptLens* leverages the principles outlined in the research paper [3], which uses symbolic Semantic Web methods to automatically induce semantic concept labels for individual neurons from a large knowledge base made from Wikipedia categories and evaluate their precision by analyzing the false positive rates of neuron activations. This approach allows users to see not only what stimuli activate specific neurons but also how confidently these neurons respond to different inputs.

2 Method

System Overview The core idea behind *ConceptLens* is to provide bar chart visualizations that contextualize the certainty of each detected concept based on the neuron activations. *ConceptLens* combines a Convolutional Neural Network (CNN) trained on specific image classes with symbolic reasoning techniques (Concept Induction) to assign semantically meaningful labels to the neurons in the final dense layer from a knowledge base of 2 million concepts. The system's backend processes images to detect concepts and calculate error-margin percentages, indicating the confidence level of each activation.

Error-Margin Analysis The core innovation of ConceptLens lies in its error-margin analysis. This measure assesses the likelihood that a given neuron activation accurately corresponds to the assigned concept by evaluating how frequently neurons activate for concepts not assigned to them on a holdout set of images. Lower error-margin percentages indicate higher confidence, while higher percentages suggest greater uncertainty. This dual focus on recall (identifying activating stimuli) and precision (evaluating responses to non-target stimuli) provides a comprehensive understanding of neuron behavior.

User Interface *ConceptLens* features a user-friendly interface that allows users to upload images and receive real-time visualizations of neuron activations. The main components of the interface include:

1. **Image Upload and Selection:** Users can upload their images or choose from a curated gallery. The tool supports a wide range of images, although results may vary for images outside the 10 classes it was primarily trained on: bathroom, bedroom, building facade, conference room, dining room, highway, kitchen, living room, skyscraper, and street.
2. **Concept Detection and Visualization:** ConceptLens processes the uploaded image through trained CNN and Concept Induction to detect relevant concepts. The detected concepts are then presented as bar chart visualization and their corresponding error-margin percentages, providing users with a clear understanding of the network's predictions.
3. **Error-Margin Display:** The interface highlights the error-margin percentages for each detected concept, allowing users to gauge the confidence of the network's predictions. Lower percentages indicate higher confidence in the concept detection.

Technical Details *ConceptLens* utilizes a ResNet50V2 architecture for its CNN, trained on a subset of the ADE20K dataset. The network's last hidden layer neurons are analyzed (see [2]) and labeled using an OWL-reasoning-based Concept Induction algorithm (ECII, [9]) over a large background knowledge base derived from Wikipedia [10]. This assigns high-level concepts to neurons, facilitating the error-margin analysis.

Error margins are calculated by evaluating neuron activations across a large dataset of images from Google and ADE20K (see [3]). This includes both target

label images (those that match the neuron's assigned concept) and non-target label images (those that do not match the concept). By comparing activation patterns, *ConceptLens* determines the likelihood of correct concept detection, thus providing valuable insights into the network's interpretability.

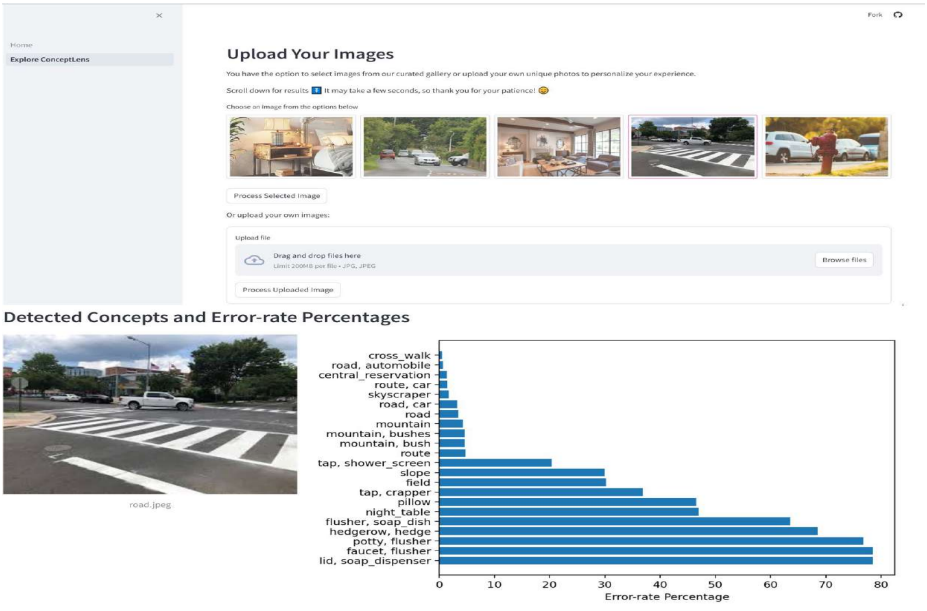

Fig. 1: ConceptLens visualizes detected concepts in images, showing their error margins. In this street scene, "cross_walk" and "road" are confidently recognized with low error percentages, while uncertainty is shown for other labels like "automobile" and "central_reservation".

Demonstration Explore ConceptLens firsthand through our interactive tool at: ConceptLens Demo.[1] Watch the demo video for a preview of its features here[2] and find the code repository on GitHub[3] for deeper exploration and implementation.

3 Conclusion

ConceptLens represents a pioneering advancement in explainable AI, offering a robust tool for visualizing and interpreting hidden neuron activations within neural networks. By integrating advanced error-margin analysis with convolutional neural networks and symbolic reasoning techniques, ConceptLens bridges critical gaps in model interpretability. Key areas for future development include: 1) Extending ConceptLens to a broader range of datasets and classes 2) Improving the user interface based on continuous user feedback 3) Developing more sophisticated error-margin analysis methodologies for deeper insights into neural network reliability.

Acknowledgement Authors acknowledge partial funding under NSF grant 2333532 *EduGate*.

References

1. Auli, M., Galley, M., Quirk, C., Zweig, G.: Joint language and translation modeling with recurrent neural networks. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1044–1054 (2013)
2. Dalal, A., Rayan, R., Barua, A., Vasserman, E.Y., Sarker, M.K., Hitzler, P.: On the value of labeled data and symbolic methods for hidden neuron activation analysis. In: 18th International Conference on Neural-Symbolic Learning and Reasoning, NeSy 2024, Barcelona, Spain. (September 2024), accepted for publication
3. Dalal, A., Rayan, R., Hitzler, P.: Error-margin analysis for hidden neuron activation labels. In: 18th International Conference on Neural-Symbolic Learning and Reasoning, NeSy 2024, Barcelona, Spain. (September 2024), accepted for publication
4. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
5. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International conference on machine learning. pp. 1764–1772. The Proceedings of Machine Learning Research (2014)
6. Hamilton, I.A.: Apple cofounder Steve Wozniak says Apple Card offered his wife a lower credit limit. Business Insider (November 2019)
7. Oikarinen, T., Weng, T.W.: CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In: International Conference on Learning Representations. ICLR (2023), https://openreview.net/forum?id=iPWiwWHc1V

[1] https://conceptlens.streamlit.app/Explore_ConceptLens
[2] https://youtu.be/yLYig1IjB9Y
[3] https://github.com/abhilekha-dalal/ConceptLens

8. Ramprasath, M., Anand, M.V., Hariharan, S.: Image classification using convolutional neural networks. International Journal of Pure and Applied Mathematics **119**(17), 1307–1319 (2018)
9. Sarker, M.K., Hitzler, P.: Efficient concept induction for description logics. In: AAAI Conference on Artificial Intelligence (2018), https://api.semanticscholar.org/CorpusID:54464695
10. Sarker, M.K., Schwartz, J., Hitzler, P., Zhou, L., Nadella, S., Minnery, B.S., Juvina, I., Raymer, M., Aue, W.: Wikipedia knowledge graph for explainable ai. In: Iberoamerican Conference on Knowledge Graphs and Semantic Web (2020), https://api.semanticscholar.org/CorpusID:225081087
11. Silberg, J., Manyika, J.: Tackling bias in artificial intelligence (and in humans). McKinsey Global Institute (June 2019)