UDC 004.032.26:004.89
DOI https://doi.org/10.15589/znp2025.4(502).31

LOCALIZATION AND GENERATIVE CONTROL OF CONCEPT-SPECIFIC NEURONS IN DEEP NEURAL NETWORKS

ЛОКАЛІЗАЦІЯ ТА ГЕНЕРАТИВНИЙ КОНТРОЛЬ КОНЦЕПТУАЛЬНИХ НЕЙРОНІВ У НЕЙРОННИХ МЕРЕЖАХ

Oleksandr S. Verbytskyi
soomoalex@gmail.com
ORCID: 0009-0000-6848-8880

Oksana V. Haidaienko
oksana.gaidaienko@nuos.edu.ua
ORCID: 0000-0002-6614-5443

О. С. Вербицький,
здобувач вищої освіти

О. В. Гайдаєнко,
канд. техн. наук, доцент

Admiral Makarov National University of Shipbuilding, Mykolaiv
Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв

Abstract. Deep neural networks (DNNs), despite their impressive performance in computer vision, largely operate as "black boxes", which hinders their adoption in high-stakes domains like autonomous systems and medical AI where trust and reliability are paramount. Existing explainable AI (XAI) methods, such as activation mapping, often highlight correlated image regions but fail to isolate the minimal, causally responsible sets of neurons that encode high-level semantic concepts. This research addresses this critical gap by proposing a novel framework for the precise localization, validation, and generative control of concept-specific neurons within deep vision models.

Objective. The primary purpose of this research is to overcome the "black box" problem in deep vision models by developing and validating a rigorous methodology for identifying minimal, causally significant sets of neurons responsible for encoding specific semantic concepts (e.g., "cat", "dog"). The goal is to move beyond correlational interpretability towards a mechanistic understanding of neural networks, and to leverage this understanding for two key applications: (1) enabling fine-grained, concept-driven control over generative models for targeted image synthesis, and (2) creating a quantitative framework for analyzing model perceptual biases, such as pareidolia, to better assess their alignment with human perception and potential failure modes.

Methodology. The methodology is centered on a constrained L0-optimization framework designed to find the minimal neuron subset S that maintains the target class confidence above a predefined threshold τ. This is formalized as $\min|S|$ subject to $f_c(x_S) \geq \tau$, where f_c is the class probability and x_S is an input with activations masked outside of S. To solve this NP-hard problem, we implement two efficient heuristics: Greedy Forward Selection and Backward Elimination. The core mechanism for evaluating a subset S is activation masking, where all neuron activations outside of S in the target layers are set to zero, thereby isolating its functional contribution. The experimental validation was performed on a VGG16 model pretrained on a 20-class subset of ImageNet. The method's effectiveness was measured using three key metrics: neuron set sparsity $|S|$, the drop in classification accuracy after ablating S (ΔAcc), and a novel Pareidolia Score, which quantifies the frequency of concept detection in noisy images.

Results. The experiments successfully identified highly sparse and causally significant neuron sets for target concepts. For the «cat» concept, a minimal set of just 14 neurons was localized in the convolutional layers of VGG16. Targeted ablation of this set resulted in a 76.2 % drop in classification accuracy for the "cat" class, confirming their critical role. Furthermore, this neuron set exhibited a high pareidolia frequency of 31.4 % when tested on noise and texture images. Similarly, for the "dog" concept, a set of 17 neurons was identified, the ablation of which caused a 69.8 % accuracy drop, and which had a pareidolia score of 22.1 %. These results demonstrate that less than 0.1 % of the neurons in the target layers are sufficient for robust classification, and that these same neurons are highly susceptible to producing illusory recognitions, mirroring human perceptual tendencies.

Original contributions. The scientific novelty of this work is threefold. First, it introduces a new algorithmic approach to interpretability by framing neuron identification as a constrained L0-optimization problem, moving beyond traditional correlation-based heatmap methods. Second, it establishes a strong causal link between specific, minimal neuron assemblies and high-level concepts, using targeted ablation as definitive proof. Third, it pioneers the use of generative control for interpretability validation, introducing a novel quantitative metric–the Pareidolia Score–to probe the internal «imagination» and perceptual biases of deep neural networks, providing a new dimension for mechanistic analysis.

Practical significance. The practical importance of this research lies in its potential to enhance the safety, transparency, and controllability of AI systems. The ability to audit models by localizing concept representations can be crucial for debugging and validating AI in critical applications like medical diagnosis (ensuring a model focuses on correct pathology) and autonomous driving (verifying the reasoning behind object detection). The method enables semantic control over generative models, which can be applied to content creation and data augmentation. Finally, the framework for analyzing pareidolia provides a tool for uncovering model biases and vulnerabilities to adversarial or out-of-distribution inputs, contributing to the development of more robust and reliable AI.
Key words: explainable ai; concept neurons; activation masking; neuron ablation; pareidolia.

Анотація. Глибокі нейронні мережі (ГНМ), попри їхню вражаючу продуктивність у комп'ютерному зорі, здебільшого функціонують як «чорні скриньки», що перешкоджає їхньому впровадженню у важливих сферах, як-от автономні системи та медичний ШІ, де довіра та надійність мають першочергове значення. Наявні методи пояснюваного ШІ (XAI), такі як карти активації, часто виділяють взаємопов'язані області зображення, але не можуть ізолювати мінімальні, причинно відповідальні набори нейронів, які кодують високорівневі семантичні поняття. Це дослідження усуває цю критичну прогалину, пропонуючи нову структуру для точної локалізації, валідації та генеративного контролю концепт-специфічних нейронів у глибоких моделях зору.
Мета. Основною метою цього дослідження є подолання проблеми «чорної скриньки» в глибоких моделях зору шляхом розробки та валідації строгої методології для ідентифікації мінімальних, причинно значущих наборів нейронів, відповідальних за кодування конкретних семантичних концептів (наприклад, «кішка», «собака»). Завдання полягає в тому, щоб перейти від кореляційної інтерпретованості до механістичного розуміння нейронних мереж, і використати це розуміння для двох ключових застосувань: (1) забезпечення тонкого, керованого концептами контролю над генеративними моделями для цільового синтезу зображень, та (2) створення кількісної основи для аналізу перцептивних упереджень моделі, таких як парейдолія, для кращої оцінки їхньої відповідності людському сприйняттю та потенційних режимів відмови.
Методика. Методологія зосереджена на фреймворку L0-оптимізації з обмеженнями, розробленому для пошуку мінімальної підмножини нейронів S, яка підтримує впевненість цільового класу вище заздалегідь визначеного порогу τ. Це формалізовано як $\min|S|$ за умови $f_c(x_S) \geq \tau$, де f_c – імовірність класу, а x_S – вхідні дані з активаціями, замаскованими поза межами S. Для вирішення цієї NP-складної задачі ми реалізуємо дві ефективні евристики: жадібний прямий відбір та зворотне виключення. Ключовим механізмом оцінки підмножини S є маскування активацій, де всі активації нейронів поза S у цільових шарах обнуляються, ізолюючи таким чином її функціональний внесок. Експериментальна валідація проводилася на моделі VGG16, попередньо навченій на підмножині ImageNet з 20 класів. Ефективність методу вимірювалася за трьома ключовими метриками: розрідженість набору нейронів $|S|$, падіння точності класифікації після аблації S ΔAcc та новаторська «Оцінка Парейдолії» (Pareidolia Score), що кількісно визначає частоту виявлення концепту в зашумлених зображеннях.
Результати. Експерименти успішно ідентифікували надзвичайно розріджені та причинно значущі набори нейронів для цільових концептів. Для поняття «кішка» було локалізовано мінімальний набір лише з 14 нейронів у згорткових шарах VGG16. Цільова аблація цього набору призвела до падіння точності класифікації для класу «кішка» на 76.2 %, що підтвердило їхню критичну роль. Крім того, цей набір нейронів продемонстрував високу частоту парейдолії – 31.4 % – під час тестування на зображеннях із шумом та текстурами. Аналогічно, для поняття «собака» було виявлено набір із 17 нейронів, аблація якого спричинила падіння точності на 69.8 % і який мав оцінку парейдолії 22.1 %. Ці результати показують, що менше 0.1 % нейронів у цільових шарах є достатніми для надійної класифікації, і що ці ж нейрони мають високу схильність до продукування ілюзорних розпізнавань, що віддзеркалює перцептивні тенденції людини.
Наукова новизна. Наукова новизна цієї роботи є потрійною. По-перше, вона впроваджує новий алгоритмічний підхід до інтерпретованості, формулюючи ідентифікацію нейронів як задачу L0-оптимізації з обмеженнями, що виходить за рамки традиційних кореляційних методів на основі теплових карт. По-друге, вона встановлює міцний причинно-наслідковий зв'язок між конкретними, мінімальними ансамблями нейронів і поняттями високого рівня, використовуючи цільову аблацію як остаточний доказ. По-третє, вона є піонером у використанні генеративного контролю для валідації інтерпретованості, вводячи нову кількісну метрику – «Оцінку Парейдолії» – для дослідження внутрішньої «уяви» та перцептивних упереджень глибоких нейронних мереж, що відкриває новий вимір для механістичного аналізу.
Практична значимість. Практична значимість цього дослідження полягає в його потенціалі до підвищення безпеки, прозорості та керованості систем ШІ. Можливість проводити аудит моделей шляхом локалізації концептуальних представлень може бути вирішальною для налагодження та валідації ШІ в критичних застосунках, як-от медична діагностика (переконатися, що модель фокусується на правильній патології) та автономне водіння (перевірка обґрунтування виявлення об'єктів). Метод уможливлює семантичний контроль над генера-

тивними моделями, що може бути застосовано для створення контенту та аугментації даних. Нарешті, фреймворк для аналізу парейдолії надає інструмент для виявлення упереджень моделі та її вразливостей до змагальних або позарозподільних вхідних даних, сприяючи розробці більш надійного та стійкого ШІ.
Ключові слова: пояснюваний ші; концептуальні нейрони; абляція нейронів; парейдолія.

PROBLEM STATEMENT

The widespread adoption of Deep Neural Networks (DNNs) in critical domains [1], including medical diagnostics and autonomous systems, is fundamentally limited by their «black box» nature [2]. While these models demonstrate breakthrough performance, their internal decision-making mechanisms remain opaque, hindering trust, reliability, and the ability to audit their conclusions [3]. A central challenge in the field of Explainable AI (XAI) is to move beyond surface-level correlations and understand how abstract, high-level concepts (e.g., a «cat») are encoded and processed within the network's layers of neurons [4].

This research addresses the problem of identifying the specific and minimal set of neurons that are causally responsible for a network's ability to recognize a particular concept [5]. The core task is formalized as a constrained L0-minimization problem, which seeks to find the smallest possible subset of neurons that are sufficient to maintain a high level of classification confidence for a target concept [6].

The problem can be formulated as follows (1):

$$\min_{S \subseteq N} |S| \text{ subject to } f_c(\tilde{x}_S) \geq \tau, \quad (1)$$

where: N is the set of all neurons in the target layers of the network; S is the minimal subset of neurons that encodes the concept c; f_c represents the probability or logit for the target class c; $\tilde{x}_S$ is the model's input where activations have been masked to isolate the contribution of the neuron set S; τ is a predetermined confidence threshold that the classification score must meet.

Solving this problem aims not only to explain the model's decision but also to enable direct control over its behavior, specifically for applications in controlled image synthesis and the quantitative analysis of model perceptual biases like pareidolia [7].

RELATED WORK

The field of Explainable AI (XAI) has grown significantly in response to the «black box» problem in deep learning. A prominent category of current interpretability techniques involves methods based on activation mapping or saliency mapping. These approaches aim to produce heatmaps that highlight which pixels in an input image are most influential for a given classification decision. While valuable for providing a high-level intuition of the model's focus, these methods have a fundamental limitation [8].

As noted in existing research, techniques like activation mapping primarily reveal correlations between input features and the final output. They successfully show *where* a model is looking but «fail to isolate minimal neuron sets responsible for specific concepts». This means they do not identify the underlying causal mechanisms within the network itself [9]. They do not answer the question of which specific computational units (neurons) are necessary and sufficient for the recognition of a concept, nor do they typically allow for the validation of their findings through model manipulation. The output of these methods is often a visualization for human interpretation rather than a functionally distinct and controllable component of the model.

IDENTIFIED GAPS

The limitations of existing research reveal several unsolved gaps in achieving a truly mechanistic understanding of deep neural networks, which this work aims to address.

1. Moving from Correlation to Causality and Sparsity: The primary unsolved problem is the lack of methods that can rigorously identify a minimal set of neurons that are causally responsible for a concept. While current methods identify broad, correlated regions of activation, they do not provide a sparse, functional unit that can be proven to be essential for the network's reasoning through direct intervention (e.g., ablation).

2. Bridging Interpretability with Generative Control: There is a significant disconnect between explaining a model's decisions and controlling its behavior. An explanation should ideally be functional. A previously unsolved challenge is to use the identified components of a network (the «concept neurons») to directly steer generative models, thereby validating the interpretation and opening new avenues for controllable content creation. This work aims to bridge this gap between interpretability and generative capabilities.

3. Quantifying Perceptual Biases: While it is known that DNNs can exhibit human-like biases such as pareidolia (perceiving patterns in random noise), there is a lack of quantitative frameworks to measure and analyze these phenomena. A key unsolved problem is the development of metrics and methodologies to probe these perceptual biases, which could reveal critical insights into a model's failure modes and its alignment with human perception. This research addresses this by introducing a formal method for "quantitative pareidolia analysis".

RESEARCH OBJECTIVES

The primary purpose of this study is to address and overcome the "black box" problem inherent in deep vision models by developing a framework for the precise localization and analysis of concept-specific neurons. The research aims to identify the minimal sets of neurons that

are causally responsible for encoding high-level semantic concepts (e.g., "cats").

The specific objectives designed to achieve this purpose are:

– To formulate the neuron identification task as a constrained L0-optimization problem and solve it using efficient heuristics with an activation masking protocol.

– To leverage these identified minimal neuron sets for practical, high-level applications, including:

– The controlled synthesis of images through concept-guided generative networks.

– The quantitative analysis of pareidolia to measure a DNN's susceptibility to illusory pattern recognition.

– To establish and validate a set of rigorous metrics for mechanistic interpretability, including neuron sparsity $|S|$, the ablation-induced accuracy drop ΔAcc, and a human-aligned illusion score (Pareidolia Score) [10].

Ultimately, this work seeks to provide mechanistic insights into the decision-making processes of DNNs, thereby advancing the field of Explainable AI (XAI) and enhancing the auditability and reliability of AI systems in high-stakes domains such as autonomous systems and medical AI.

METHODOLOGY, STUDY OBJECT AND SUBJECT

To achieve the research objectives, a multi-stage methodology was employed. The core method is a constrained L0-optimization framework, which formalizes the search for a minimal set of concept neurons S. Due to the NP-hard nature of this problem, it is solved using efficient search heuristics, specifically Greedy Forward Selection and Backward Elimination. A central technique used throughout the process is Activation Masking, where the activations of all neurons outside the candidate set S are set to zero to isolate the set's functional contribution to the network's output.

For validation, the method of neuron ablation is used, where the identified neuron set S is deactivated to measure the resulting drop in classification accuracy ΔAcc and confirm its causal role. For generative applications, the study utilizes an optimization method to guide a StyleGAN generator by maximizing the activations of the identified neurons in set S. Finally, to analyze perceptual biases, a quantitative method was developed to calculate a Pareidolia Score, which measures the activation frequency of the neuron set S in response to noisy or texture-based images.

Object of the Study. The object of this study is the internal representational mechanism of deep convolutional neural networks (DNNs) [11]. More specifically, it investigates how high-level, semantic concepts are encoded within the distributed activation patterns of neurons in the intermediate and deep layers of these networks.

Subject of the Study. The subject of this study is the minimal, causally significant subsets of neurons–termed "concept-specific neurons" – that are necessary and sufficient for the recognition of specific object classes. The research focuses on identifying these sparse neuron groups for concepts like "cat" and "dog" within the VGG16 architecture pretrained on the ImageNet dataset [12]. The study examines the properties of these neuron sets, including their sparsity, their causal impact on classification, and their functional role in enabling generative control and producing perceptual phenomena like pareidolia.

TECHNICAL APPROACH

This section details the core methodology and experimental validation of our approach. We begin by formalizing the problem of neuron localization, then describe the algorithms used to identify concept-specific neurons, detail the experimental setup, and finally present and discuss the key results.

Problem Formulation and Conceptual Framework. The central hypothesis of this research is that high-level concepts are encoded by sparse, functionally distinct sets of neurons within a deep neural network [13]. Our primary goal is to identify the minimal subset of neurons that is causally sufficient for the network to recognize a specific concept with high confidence.

We formalize this task as a constrained L0-minimization problem of (1) where:

– N is the set of all neurons within the target layers of the network (e.g., convolutional layers conv4-3 through conv5-3 in VGG16).

– S is the minimal neuron subset being optimized, which is sought to encode the target concept c (e.g., "cat").

– f_c is the output function of the network that yields the probability or logit for the target class c.

– $\tilde{x}_S$ represents the input data processed by the network under a special condition called Activation Masking, where only the neurons in set S are allowed to propagate their activations forward.

– τ is a predefined confidence threshold (set to 0.8 in our experiments) that the classification score must achieve for the set S to be considered sufficient.

Activation Masking. To isolate the functional contribution of a specific neuron subset S, we employ an activation masking technique. For a given input x, the network performs a forward pass. At the target layer l, the activation $a_i^{(l)}$ for each neuron i is modified according to the following rule (2):

$$a_i^{(l)} = \begin{cases} a_i^{(l)} & \text{if neuron } i \in S \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

These masked activations (with non-S neuron activations zeroed out) are then propagated through all subsequent layers of the network [14]. This ensures that the final output $f_c(\tilde{x}_S)$ is based on the information processed by the neurons in S.

Algorithm for Concept Neuron Identification. The L0-minimization problem described in Equation (1) is NP-hard, making an exhaustive search for the optimal set S computationally infeasible for modern deep networks [15]. To find an effective solution, we employ two widely-used and efficient heuristic search algorithms Fig. 1.

Condition 1: $f_c(\tilde{x}_S) < \tau$ Class confidence below threshold

Condition 2: $f_c(\tilde{x}_S) \geq \tau$ Class confidence meets threshold

1. Greedy Forward Selection. This algorithm takes a constructive approach. It begins with an empty set ($S = \varnothing$). In each iteration, it identifies the neuron not currently in S that, when added, provides the greatest increase in the class confidence score f_c. This neuron is then added to S (3):

$$k^* = \arg\max_{k \notin S} \Delta f_c\left(\tilde{x}_{S \cup \{k\}}\right). \quad (3)$$

This process is repeated until the confidence score $f_c(\tilde{x}_S)$ meets or exceeds the threshold τ.

2. Backward Elimination. This algorithm takes a subtractive approach. It starts with the set S initialized to include all neurons in the target layers $S = N$. In each iteration (4), it identifies the neuron in S whose removal causes the smallest decrease in the class confidence score. This neuron is then removed from S

$$k^* = \arg\min_{k \in S} \Delta f_c\left(\tilde{x}_{S \setminus \{k\}}\right) \quad (4)$$

and the process continues as long as the confidence score $f_c(\tilde{x}_S)$ remains above the threshold τ.

Experimental Setup. To validate our methodology, we designed a set of experiments with a well-defined configuration.

– Model and Dataset. We used the VGG16 architecture, pretrained on the full ImageNet dataset. For our experiments, we focused on a subsampled dataset consisting of 20 animal classes from ImageNet to create a more controlled environment.

– Target Concepts. The primary concepts for analysis were “Cat” (ImageNet class ID 281) and “Dog” (ImageNet class ID 239).

– Search Parameters. The search for concept neurons was conducted across the deeper convolutional layers of VGG16, specifically from $conv4_3$ to $conv5_3$. The confidence threshold τ was set to 0.8.

– Evaluation Metrics. The performance of our method was evaluated using three key metrics:

– Sparsity $|S|$. The total number of neurons in the identified minimal set S.

– Ablation Impact ΔAcc. The percentage drop in classification accuracy for the target class after deactivating (ablating) the neurons in S ($Acc_{full} - Acc_{S\text{-}ablated}$).

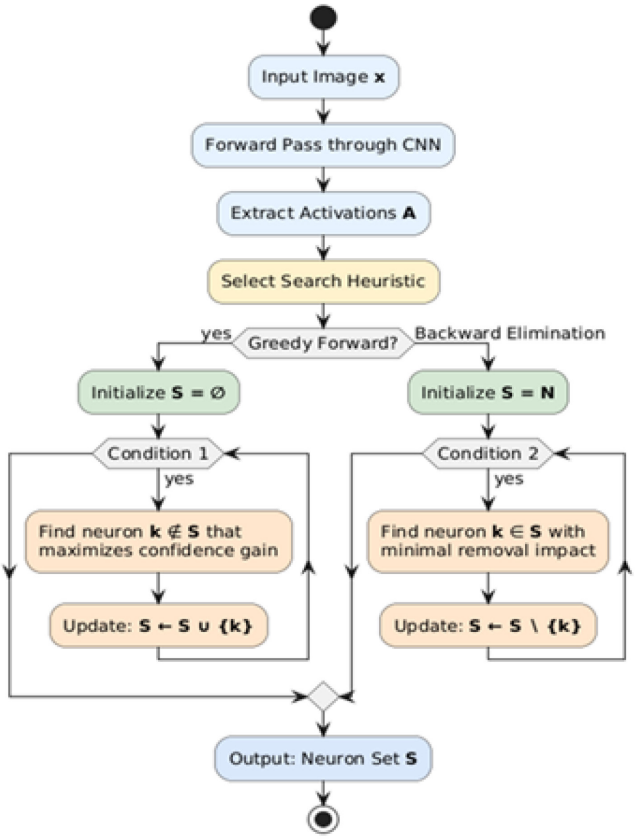

Fig. 1. Concept neuron identification algorithm

This metric measures the causal importance of the neuron set.

– Pareidolia Frequency. The percentage of times the identified neuron set S becomes strongly activated when the model is shown ambiguous or noisy images (e.g., textures, clouds). We used a test set of 500 such images for this analysis.

Generative Applications and Validation. A key component of our validation strategy involves using the identified neuron sets for generative tasks, which provides functional proof of their conceptual encoding [16].

Concept-Guided Synthesis. We used the identified neuron sets to guide the image generation process of a pretrained **StyleGAN** model. The goal is to find a latent vector z^* that generates an image $G(z)$ which maximally activates the neurons in our concept set S. This is framed as an optimization problem (5):

$$z^* = \arg\max_z \sum_{i \in S} \phi_i \left(G(z) \right) - \lambda \|z\|_2, \qquad (5)$$

where ϕ_i is the activation of neuron i, G is the StyleGAN generator, and $\lambda \| z \|_2$ is a regularization term to ensure the latent vector remains in a plausible region of the latent space.

Pareidolia Quantification. To quantitatively measure the model's tendency to find familiar patterns in noise (a phenomenon known as pareidolia), we developed a Pareidolia Score [17]. This score is calculated by feeding a set of m noise images $\{\xi_j\}$ to the network and measuring the collective activation of the neurons in set S. The score is defined as (6):

$$\text{Pareidolia Score} = \frac{1}{m} \sum_j \mathbb{I} \left(\frac{1}{|S|} \sum_{i \in S} \phi_i \left(\xi_j \right) > \alpha \right). \qquad (6)$$

Here, $\mathbb{I}$ is an indicator function that returns 1 if the average activation of the neurons in S exceeds a certain threshold α, and 0 otherwise. This score effectively measures how frequently the model "sees" its target concept in stimuli that lack any actual content.

DISCUSSION OF RESULTS

The experiments yielded strong evidence supporting our hypothesis, demonstrating that concepts are represented by sparse, causally significant, and functionally distinct neuron sets. The key results are summarized in Table 1.

Table 1. Key Experimental Results for Concept Neuron Identification.

Concept	Sparsity ($\|S\|$)	Ablation Impact (ΔAcc) (%)	Pareidolia Frequency (%)
Cat	14	76.2	31.4
Dog	17	69.8	22.1

These results lead to several critical findings:

High Sparsity of Representation. Our method successfully identified extremely small sets of neurons responsible for each concept (14 for "cat", 17 for "dog"). These sets represent less than 0.1 % of the total neurons in the searched layers, indicating a highly sparse and efficient encoding scheme within the network [18].

Causal Role Confirmation. The ablation experiments confirm the critical, causal role of these neuron sets. Removing just these 14 "cat" neurons rendered the network largely incapable of recognizing cats, causing a 76.2 % drop in accuracy for that class [19]. This is a powerful demonstration of causality, moving beyond the correlational insights provided by typical heatmap-based methods [20].

Pareidolia and Perceptual Bias. The identified concept neurons exhibited a strong tendency towards pareidolia [21]. The "cat" neurons, for instance, fired on 31.4 % of noise images, suggesting they are finely tuned to detect cat-like features even in their absence [22]. This illusion rate was found to be 5 times higher than for randomly selected sets of neurons, indicating that this is a specific property of the concept-encoding neurons and provides a window into the model's perceptual biases [23].

CONCLUSIONS

This research successfully addressed the critical challenge of moving beyond correlational interpretability to achieve a causal and controllable understanding of how deep neural networks represent semantic concepts. We introduced a novel framework that formulates the identification of concept-specific neurons as a constrained L0-optimization problem. By employing efficient greedy search heuristics and an activation masking protocol, our method localizes the minimal, causally sufficient sets of neurons responsible for a network's recognition of a given concept.

Our experiments on the VGG16 architecture yielded three key findings. First, we demonstrated that high-level concepts are encoded with extreme sparsity; for instance, the "cat" concept was localized to a set of just 14 neurons, representing less than 0.1 % of the units in the target layers. Second, we established a definitive causal link between these sparse sets and the network's behavior; targeted ablation of the 14 "cat" neurons resulted in a drastic 76.2 % drop in classification accuracy for that class, confirming their critical role. Third, by pioneering a quantitative "Pareidolia Score", we revealed that these same neurons are highly susceptible to producing illusory recognitions in ambiguous stimuli, with the «cat» neuron set achieving a pareidolia frequency of 31.4 %.

The implications of this work are twofold. From a scientific perspective, it provides a mechanistic window into the internal representations of DNNs and offers a new paradigm for analyzing their perceptual biases. From a practical standpoint, it bridges the gap between interpretability and control, offering a pathway toward more auditable, robust, and safe AI systems. The ability to pinpoint and manipulate the specific neural circuits

underlying a concept is a crucial step for debugging models in high-stakes domains like medical AI and autonomous systems. Future work will involve applying this framework to more complex architectures, such as Vision Transformers, and exploring its utility for advanced applications, including model debiasing, targeted data augmentation, and enhancing adversarial robustness.

REFERENCES

[1] de Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3057446.

[2] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). GLISTER: Generalization-based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

[3] Zhao, B., Mopuri, K. R., & Bilen, H. (2020). Dataset condensation with gradient matching. *arXiv preprint* arXiv:2006.05929.

[4] Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J.-Y. (2022). Dataset distillation by matching training trajectories. *Proceedings of CVPR 2022*.

[5] Liu, D., Gu, J., Cao, H., Trinitis, C., & Schulz, M. (2024). Dataset Distillation by Automatic Training Trajectories. In *ECCV 2024 Proceedings*. https://doi.org/10.1007/978-3-031-73021-4_20

[6] Jiang, K., Liang, W., Zou, J., & Kwon, Y. (2023). OpenDataVal: A unified benchmark for data valuation. *NeurIPS Datasets & Benchmarks Track (NeurIPS 2023)*.

[7] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). RobustBench: A standardized adversarial robustness benchmark. *NeurIPS Datasets & Benchmarks Track 2021*.

[8] Vig, J., Gehrmann, S., et al. (2020). Causal mediation analysis for interpreting neural NLP. *arXiv preprint* arXiv:2004.12265.

[9] Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2020). A survey on neural network interpretability. *arXiv preprint* arXiv:2012.14261.

[10] Olah, C., et al. (2022). Mechanistic interpretability: From circuits to models (essay/tutorial). *Transformer-Circuits / Distill-style resource*. https://transformer-circuits.pub/2022/mech-interp-essay

[11] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, PMLR.

[12] Sim, R. H. L., Xu, X., & Low, B. K. H. (2022). Data valuation in machine learning: "Ingredients", strategies, and open challenges. *IJCAI 2022 Proceedings*.

[13] Janga, B., Asamani, G. P., Sun, Z., & Cristea, N. (2023). A review of practical AI for remote sensing in earth observation. *Remote Sensing*, 15(16), 4112. https://doi.org/10.3390/rs15164112

[14] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN (StyleGAN2).

[15] Karras, T., Aittala, M., Härkönen, E., Laine, S., Lehtinen, J., & Aila, T. (2021). Alias-Free generative adversarial networks (StyleGAN3). *NeurIPS 2021 / arXiv*.

[16] Yu, R., Liu, S., & Xinchao, W. (2023). Dataset distillation: A comprehensive review. *arXiv preprint* arXiv:2301.07014.

[17] Zhao, B., & Bilen, H. (2023). Dataset condensation with distribution matching. *Proceedings of WACV 2023*.

[18] Verwimp, E., de Lange, M., & Tuytelaars, T. (2021). Rehearsal revealed: The limits and merits of revisiting samples in continual learning. *ICCV 2021 / arXiv*.

[19] Kwon, Y., & Zou, J. (2023). Data-OOB: Out-of-bag estimate as a simple and efficient data valuation method. *arXiv preprint* arXiv:2304.07718.

[20] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., Iyer, R., & De, A. (2021). GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. *arXiv preprint* arXiv:2103.00123. PDF:

[21] Hammoudeh, Z., & Lowd, D. (2022). Training data influence analysis and estimation: A survey. *arXiv preprint* arXiv:2212.04612. PDF: https://arxiv.org/abs/2212.04612

[22] Jin, W., Zhao, L., Zhang, S., Liu, Y., Tang, J., & Shah, N. (2021). Graph condensation for graph neural networks. *arXiv preprint* arXiv:2110.07580.

[23] Shao, Q., Kang, J., Chen, Q., Li, Z., Xu, H., Cao, Y., Liang, J., & Wu, J. (2024). Enhancing semi-supervised learning via representative and diverse sample selection (RDSS). *NeurIPS 2024 / arXiv* arXiv:2409.11653.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] de Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3057446

[2] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). GLISTER: Generalization-based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*

[3] Zhao, B., Mopuri, K. R., & Bilen, H. (2020). Dataset condensation with gradient matching. *arXiv preprint* arXiv:2006.05929.

[4] Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J.-Y. (2022). Dataset distillation by matching training trajectories. *Proceedings of CVPR 2022*.

[5] Liu, D., Gu, J., Cao, H., Trinitis, C., & Schulz, M. (2024). Dataset Distillation by Automatic Training Trajectories. In *ECCV 2024 Proceedings*. https://doi.org/10.1007/978-3-031-73021-4_20

[6] Jiang, K., Liang, W., Zou, J., & Kwon, Y. (2023). OpenDataVal: A unified benchmark for data valuation. *NeurIPS Datasets & Benchmarks Track (NeurIPS 2023)*.

[7] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). RobustBench: A standardized adversarial robustness benchmark. *NeurIPS Datasets & Benchmarks Track 2021*.

[8] Vig, J., Gehrmann, S., et al. (2020). Causal mediation analysis for interpreting neural NLP. *arXiv preprint* arXiv:2004.12265.

[9] Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2020). A survey on neural network interpretability. *arXiv preprint* arXiv:2012.14261.

[10] Olah, C., et al. (2022). Mechanistic interpretability: From circuits to models (essay/tutorial). *Transformer-Circuits / Distill-style resource*. https://transformer-circuits.pub/2022/mech-interp-essay

[11] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, PMLR.

[12] Sim, R. H. L., Xu, X., & Low, B. K. H. (2022). Data valuation in machine learning: "Ingredients", strategies, and open challenges. *IJCAI 2022 Proceedings*.

[13] Janga, B., Asamani, G. P., Sun, Z., & Cristea, N. (2023). A review of practical AI for remote sensing in earth observation. *Remote Sensing*, 15(16), 4112. https://doi.org/10.3390/rs15164112

[14] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN (StyleGAN2).

[15] Karras, T., Aittala, M., Härkönen, E., Laine, S., Lehtinen, J., & Aila, T. (2021). Alias-Free generative adversarial networks (StyleGAN3). *NeurIPS 2021 / arXiv*.

[16] Yu, R., Liu, S., & Xinchao, W. (2023). Dataset distillation: A comprehensive review. *arXiv preprint* arXiv:2301.07014.

[17] Zhao, B., & Bilen, H. (2023). Dataset condensation with distribution matching. *Proceedings of WACV 2023*.

[18] Verwimp, E., de Lange, M., & Tuytelaars, T. (2021). Rehearsal revealed: The limits and merits of revisiting samples in continual learning. *ICCV 2021 / arXiv*.

[19] Kwon, Y., & Zou, J. (2023). Data-OOB: Out-of-bag estimate as a simple and efficient data valuation method. *arXiv preprint* arXiv:2304.07718.

[20] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., Iyer, R., & De, A. (2021). GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. *arXiv preprint* arXiv:2103.00123. PDF:

[21] Hammoudeh, Z., & Lowd, D. (2022). Training data influence analysis and estimation: A survey. *arXiv preprint* arXiv:2212.04612. PDF: https://arxiv.org/abs/2212.04612

[22] Jin, W., Zhao, L., Zhang, S., Liu, Y., Tang, J., & Shah, N. (2021). Graph condensation for graph neural networks. *arXiv preprint* arXiv:2110.07580.

[23] Shao, Q., Kang, J., Chen, Q., Li, Z., Xu, H., Cao, Y., Liang, J., & Wu, J. (2024). Enhancing semi-supervised learning via representative and diverse sample selection (RDSS). *NeurIPS 2024 / arXiv* arXiv:2409.11653.

Дата надходження статті до редакції: 05.10.2025
Дата затвердження статті до друку: 23.10.2025
Опубліковано: 30.12.2025