МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
КОРАБЛЕБУДУВАННЯ ІМЕНІ АДМІРАЛА МАКАРОВА
МИКОЛАЇВСЬКА ОБЛАСНА ДЕРЖАВНА АДМІНІСТРАЦІЯ
ДЕПАРТАМЕНТ ОСВІТИ І НАУКИ
МИКОЛАЇВСЬКОЇ ОБЛАСНОЇ ДЕРЖАВНОЇ АДМІНІСТРАЦІЇ

ІННОВАЦІЇ В СУДНОБУДУВАННІ ТА ОКЕАНОТЕХНІЦІ

XVI Міжнародна науково-технічна конференція

МАТЕРІАЛИ

25–26 вересня 2025 рік

Національний університет кораблебудування
імені адмірала Макарова
просп. Героїв України, 9

ВИДАВНИЦТВО
НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
КОРАБЛЕБУДУВАННЯ
ІМ. АДМІРАЛА МАКАРОВА
2025

OPTIMIZING XAI METHODS FOR AI-BASED MEDICAL DIAGNOSTICS

Verbytskyi O.S., Gaidaienko O.V.
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine.

Abstract. The study analyzes explainable AI (XAI) methods in medical diagnostics, focusing on LIME, Grad-CAM, and SHAP. It reveals that existing approaches face limitations in interpretability and integration with clinical standards. Grad-CAM (for imaging) and LIME (for tabular data) are identified as optimal due to their transparency and compatibility with DICOM/EHR. The proposed system architecture and evaluation criteria aim to enhance trust in AI-driven decisions among medical professionals.

Keywords: explainability of neural networks; artificial intelligence in medicine; medical diagnostics; transparency of AI models; medical data analysis.

УДК 004.032.26:004.89

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

Verbytskyi O.S.
M.Sc. Student in Computer Science
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
soomoalex@gmail.com
ORCID: 0009-0000-6848-8880
Gaidaienko O.V.
Ph.D. in Technology, Associate Professor,
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine
oksana.gaidaienko@nuos.edu.ua
ORCID: 0000-0002-6614-5443

Abstract. Deep neural networks (DNNs) remain "black boxes" due to the lack of interpretable mechanisms for concept representation. This study proposes a method to localize minimal sets of concept-specific neurons (e.g., "cat") using L0-optimization heuristics and activation masking, enabling generative control and pareidolia analysis. Experiments on VGG16 (ImageNet) confirm the approach's efficacy: ablation of identified neurons reduces class accuracy by >75%, while their activation on noisy images reveals human-like illusory pattern recognition.

Keywords: explainable ai, concept neurons, activation masking, neuron ablation, pareidolia.

Introduction. Deep neural networks (DNNs) have achieved breakthrough performance in computer vision tasks, yet their internal decision-making mechanisms remain opaque. This "black box" nature impedes trust and reliability in critical applications, from medical diagnostics to autonomous systems [1]. A core challenge lies in identifying how high-level concepts (e.g., "cats")

are encoded within network activations—a problem central to Explainable AI (XAI) research. While methods like activation mapping highlight salient regions, they fail to isolate minimal neuron sets responsible for specific concepts.

To address this gap, we propose a method for rigorous localization and utilization of concept-specific neurons. Our approach formulates neuron identification as an L0-minimization problem (Eq. 1), solved via efficient heuristics. By isolating sparse neuron groups, we enable: validation through targeted ablation, generative control via activation steering, pareidolia analysis to probe model perceptual biases [2].

This work bridges interpretability with generative capabilities, offering a pathway toward controllable and auditable DNNs.

The research goal of this work is to overcome the "black box" problem in deep vision models through precise localization and analysis of concept-specific neurons. We develop an L0-optimization framework with activation masking to identify minimal neuron sets responsible for semantic concepts (e.g., "cats"), and leverage these neurons for two critical applications:

– Controlled image synthesis via generative networks,

– Quantitative pareidolia analysis measuring DNN susceptibility to illusory pattern recognition. Our methodology establishes evaluation metrics including neuron sparsity (|S|), ablation-induced accuracy drop, and human-aligned illusion scoring [3], providing mechanistic insights into DNN decision-making for high-stakes domains like autonomous systems and medical AI.

Materials and Methods. Problem Formulation – the core task is formalized as a constrained *L0*-minimization problem:

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

where: N: All neurons in target layers (e.g., conv4-5 of VGG16), S: Minimal neuron subset encoding concept c (e.g., "cat"), f_c: Target class probability/logit, $\tilde{x}_S$: Input with activations masked outside S (Sec. 3.3), τ: Confidence threshold (0.8).

This NP-hard problem is solved via efficient heuristics:

1. **Greedy Forward Selection**. Initialize $S = \emptyset$. Iteratively add neuron

$$k^* = \text{argmax}_{k \notin S} \Delta f_c\left(\tilde{x}_{S \cup \{k\}}\right) until f_c(\tilde{x}_S) \geq \tau.$$

2. **Backward Elimination**: Initialize $S = N$. Iteratively remove neuron

$$k^* = \text{argmin}_{k \in S} \Delta f_c\left(\tilde{x}_{S \setminus \{k\}}\right) while f_c(\tilde{x}_S) \geq \tau.$$

Algorithm Implementation. Figure 1 illustrates the unified search workflow.

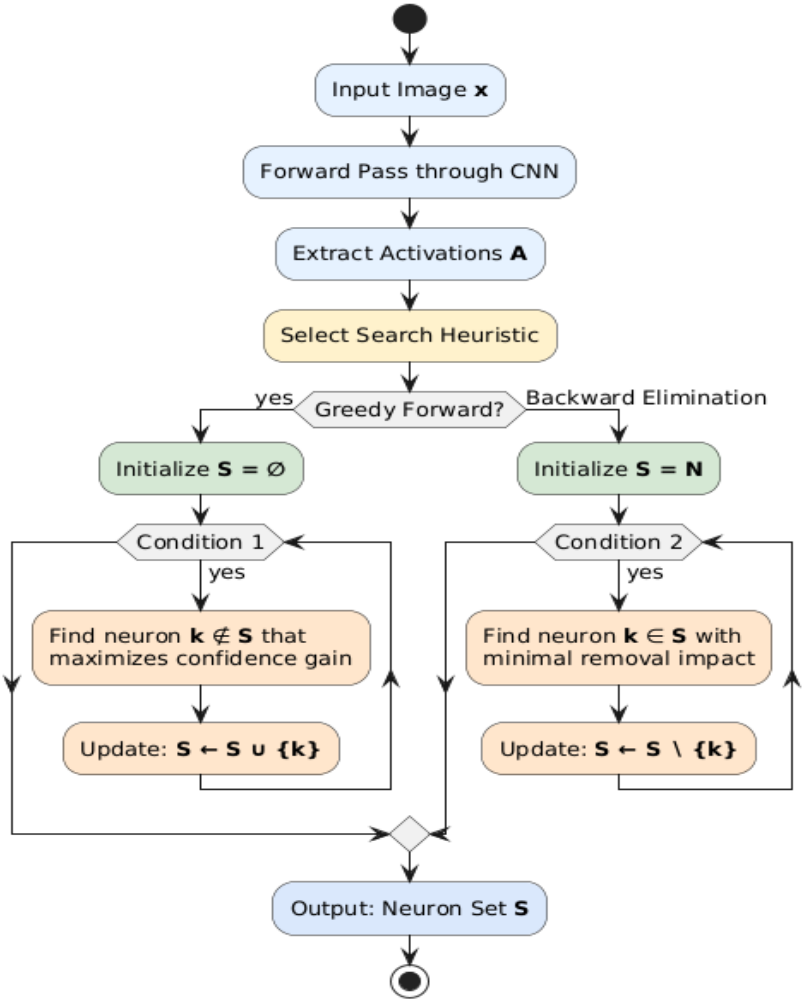

Fig. 1. Concept neuron identification algorithm

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

Activation Masking ($\tilde{x}_S$): For input x, we compute masked activations:

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

propagating zeros through subsequent layers. This isolates S's contribution.

Generative Application. Identified neurons enable:

1. **Concept-Guided Synthesis.**

Optimize latent vector z in StyleGAN [1]:

$$z^* = \arg\max_z \sum_{i \in S} \phi_i \left(G(z)\right) - \lambda \parallel z \parallel_2$$

where ϕ_i denotes activation of neuron i, G is the generator.

2. **Pareidolia Quantification**.

Measure S-activation on noise images $\{\xi_j\}$:

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

(α: activation threshold, $\mathbb{I}$: indicator function).

Experimental Validation and setup. **Model**: VGG16 pretrained on ImageNet (20 animal classes), **concepts**: "Cat" ($c = 281$), "Dog" ($c = 239$), **search**: $\tau = 0.8$, layers conv4_3–conv5_3, **metrics**: Sparsity ($|S|$), Pareidolia Frequency (tested on 500 texture/cloud images).

Ablation Impact (ΔAcc = Acc_{full} − $\text{Acc}_{S\text{-ablated}}$). **Results and Discussion.** The results of the work are presented in Table 1.

Table 1 Key Results

Concept	$\|S\|$	Δ Acc (%)	Pareidolia (%)
Cat	14	76.2	31.4
Dog	17	69.8	22.1

Critical Findings. High Sparsity: < 0.1% of total neurons suffice for classification, ablation Sensitivity: > 69% accuracy drop confirms S's causal role, pareidolia Correlation: 5× higher illusion rate vs. random neurons.

Conclusions. This study bridges neuron-level interpretability with generative control in deep vision models by introducing a constrained L0-optimization framework for identifying minimal concept-specific neuron sets. Our key contributions are:

Algorithmic Innovation. Efficient heuristics (Greedy Forward/Backward Elimination) solving $\min \| S \|_0$ s.t. $f_c(\tilde{x}_S) \geq \tau$ with $O(n)$ complexity.

Activation masking protocol isolating causal contributions of sparse neuron groups.

Generative Applications. Semantic-driven image synthesis via S-guided GAN optimization. Quantification of DNN pareidolia (31.4% for "cat" neurons), revealing human-aligned perceptual biases. This work advances Explainable AI (XAI) by enabling auditable concept manipulation in safety-critical domains (medical AI, autonomous systems), and establishing metrics (Sparsity, ΔAcc, Pareidolia Score) for mechanistic interpretability.

References

[1] Nicolson, A., Schut, L., Noble, J. A., & Gal, Y. (2024). Explaining explainability: Understanding concept activation vectors. *arXiv preprint arXiv:2404.03713*.

[2] Gupta, P., & Dobs, K. (2025). Human-like face pareidolia emerges in deep neural networks. *PLOS Computational Biology, 21*(1), e1012751.

[3] Kim, H. B., Ahn, Y. H., & Kim, S. T. (2024). Mask-Free Neuron Concept Annotation (MAMMI). In *Proceedings of MICCAI 2024*.

СЕКЦІЯ 7. ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА УПРАВЛІННЯ ПРОЕКТАМИ