sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 3
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0020

DeConvolve: Towards Textually Explainable and Human Cognizable Convolutional Networks

Meeradevi[1], *Hrishikesh Haritas*[1], *Darshan Bankapure*[1], *Divyansh Mishra*[2]

[1]*Department of Artificial Intelligence and Machine Learning, Ramaiah Institute of Technology, Bengaluru, India*
[2]*Department of Artificial Intelligence and Data Science, Ramaiah Institute of Technology, Bengaluru, India*
E-mails: meera_ak@msrit.edu (*corresponding author*) *1ms21ai022@msrit.edu 1ms21ai016@msrit.edu 1ms21ad022@msrit.edu*

Abstract**:** *Convolutional Neural Networks* (*CNNs*) *have demonstrated remarkable accuracy and are employed in different applications. However, adding existing CNNs to physics-aware frameworks can distort image features, reducing classification accuracy. To overcome this, a new term is added to the loss function to reduce distortions and highlight human-recognizable structures in the feature maps. The proposed DeConvolve is an explainability methodology for multimodal Large Language Models* (*LLM*) *on feature maps to extract human-understandable sub-steps and provide textual explanations for model inference. DeConvolve recognizes three major impediments when using LLMs to describe feature maps: scattered regions of interest within the feature map, large areas of interest, and conflicting learning across filters in each convolutional layer. Finally, explanations for specific toy examples are derived through weighted semantic averaging. The data is curated in the format of images, classes, and the rationale behind a professional's classification to train a Contrastive Language–Image Pre-training* (*CLIP*)*-based model for generating robust explanations.*

Keywords**:** *Convolutional neural networks, Contrastive language-image pretraining, Gradient-weighted class activation mapping, Human recognizable, Large language models.*

1. Introduction

The success of Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), has transformed the fields of medical image analysis and industrial automation [1, 2]. However, early detection and proper treatment of images can help reduce these overwhelming statistics [3]. Regression models offer the advantage of low complexity but are limited in terms of explainability. DL with neural networks takes full advantage of a dataset's complexity and enables the

redevelopment of data used in decision-making through explainable artificial intelligence approaches [4, 5]. Nevertheless, determining these characteristics requires more than just automatic human perception. The need for domain knowledge makes this approach less user-friendly for non-experts compared to visual data-related operations [6-7]. Furthermore, the required radiation dose is not high, unlike traditional imaging approaches, and the produced images exhibit superior resolution and contrast [8-9]. Therefore, a stable and cost-effective method is sought [10-11]. The proposed method is hypothesized by combining deep learning with eXplainable Convolution Neural Network (XCNN) algorithms that produce high-performance, interpretable models suitable for clinical application in the automated detection of images. Designing explainable decisions in DL-based systems is crucial to improving the trust of physicians in the system and promoting its application in the field [12-14]. Explainable classifications with greater accuracy denote a significant step toward enhancing the trust of patients and caregivers in computer-aided diagnostic systems [15-17]. The capabilities of DL approaches have advanced classification and detection to a further state. As a result, modern image processing techniques have enhanced feature learning [18-19]. While many studies focus on individual algorithms, the aim is to consolidate various algorithms and demonstrate their effectiveness with high accuracy [20]. There are two mainstream methods for enhancing the interpretability of neural networks: providing local XCNN and global XCNN for existing neural networks [21-22]. Initially, an idealized attribution benchmark dataset is considered, where a CNN is trained to classify images of circular and square frames based on the area occupied by each frame type [23-25]. This research aims to develop a methodology for XCNNs in computer vision tasks. It seeks to generate coherent, human-understandable textual explanations of a CNN's reasoning by leveraging multimodal LLMs. The goal is to create a comprehensive framework for practical deployment, fostering responsible AI adoption in critical domains such as healthcare, finance, and legal proceedings, thus enhancing the use of AI in high-stakes decision-making scenarios [26].

In the current scenario, CNNs are applied for image recognition, object detection, and medical image analysis to capture spatial patterns through convolutional layers efficiently. These layers use learnable filters inspired by the visual cortex, enabling CNNs to outperform traditional methods. Huang et al. [27] introduced DNN to enhance the explainability and physics-awareness of the DL technique. The CNN approach improved classification performance in limited labelled data using the counterpart data of the model. However, CNN was integrated into the framework by leveraging the physics-aware feature of the image, which negatively affected the classification accuracy. Shajalal, Boden and Stevens [28] developed an Explainable Artificial Intelligence (XAI) model mainly concentrated on describing networks to specialists, aiming to make image recognition human-understandable. This involved interpreting and explaining a predictive system to determine the most significant attributes, helping to better understand a network's decision-making prime concerns. However, the CNNs employed in predictive model decision-making introduce various features in a way that influences the predictions of the models. Dasari and Bhukya [29]

implemented a DNN model for automatic extraction of the classification features, incorporating an explainable method into the CNN model. The CNN-LSTM-based method, namely EdeepVPP, was an interpretable CNN approach designed for pattern extraction and probability generation to ensure enhanced sequence classification performance. The EdeepVPP faced challenges with simple array operations, resulting in lower computational complexity than the Conv2D. Luo et al. [30] presented a CNN-based selective fixed-filter active noise control approach (SFANC) using a pre-trained model. The selected control filter was delivered to a time controller operating at a parallel sampling rate, enabling reduced noise delay in the CNN-based SFANC approach. However, the CNN-assisted SFANC approach faced challenges, including a limited, slow convergence rate, poor tracking capability, and greater potential for divergence. Begum et al. [31] introduced a Lightweight CNN (LCNN) combined with the Long Short-Term Memory (LSTM) technique to improve defect prediction accuracy. The eXplainable Artificial Intelligence (XAI) approach was involved in developing deep models that efficiently managed defect prediction while enhancing performance. The main aim of the LCNN model was to enhance its ability to identify software defect features. However, image recognition within this model was limited by its adaptability to software practices. Incorporating the existing CNN into a physics-aware framework distorts image features, which negatively impacts classification accuracy by altering the image's inherent physical properties. Therefore, this research proposes the novel DeConvolve explainability methodology, which demonstrates the significant potential of using multimodal Large Language Models (LLMs) on feature maps to extract human-understandable sub-steps and provide textual explanations for model inference.

The main contributions of this research are noted below:

- The CNN efficiently handles complex patterns and decision-making with human-recognizable patterns and textual explanations, thereby enhancing the AI system.
- The integration of LLMs in multimodal AI systems demonstrates how language-based reasoning augments the interpretability of vision-based models like CNNs.
- Grad-CAM optimizes heatmaps for better alignment with the LLM process, generating filter heatmaps that are processed to highlight critical regions in a format conducive to LLM-based reasoning.

The remainder of the paper is organized as follows: Section 2 reviews related work and the state of the art, motivating the need for explainable CNNs through a survey of existing methods, Section 3 details the proposed methodology and its functioning, while Section 4 discusses the experimental results and discussion, and finally, Section 5 concludes the research.

2. Related work and state of the art

Recent advancements in Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), have significantly improved performance across domains such

as medical imaging and industrial automation [1, 2]. Despite the success of CNNs in enhancing detection accuracy, their black-box nature has raised concerns about trust and interpretability, especially in high-stakes applications [3, 4]. Regression-based models offer simplicity and computational efficiency but are limited in their capacity to model complex feature hierarchies or provide transparent reasoning [5]. In contrast, DL approaches effectively exploit high-dimensional data and have been extended with eXplainable Artificial Intelligence (XAI) methods to enhance interpretability [6, 7]. However, CNNs often require domain-specific insights to yield meaningful interpretations, which can hinder their adoption by non-experts [8, 9]. While imaging technologies with low radiation exposure and high resolution offer tangible benefits [10], translating these advancements into clinically actionable and trustworthy AI systems remains a challenge [11]. Consequently, hybrid frameworks combining interpretability and performance – such as eXplainable CNNs (XCNNs) – are increasingly investigated to bridge this gap [12, 13]. These approaches aim to build transparent decision pipelines that align with clinician reasoning and foster trust in model outputs [14, 15].

Explainability is not only a technical challenge but also a socio-technical one, where building trust in automated systems requires models to produce decisions that are not only accurate but also comprehensible [16, 17]. DL has empowered modern image processing systems to autonomously learn abstract features, but the interpretability of these features remains limited [18, 19]. Many existing studies address individual techniques without presenting a unified approach for integrating multiple interpretability mechanisms into robust predictive pipelines [20]. Two primary strategies have emerged to enhance CNN interpretability: local explanations, which highlight specific regions influencing a single prediction, and global explanations, which describe model behavior over the entire dataset [21, 22]. Early benchmark experiments used geometric primitives like circles and squares to train CNNs, serving as a foundation for building attribution models that evaluate the model's attention and feature extraction mechanisms [23-25]. Such studies laid the groundwork for developing structured methodologies for explainable CNNs in real-world applications. In the domain of multimodal explainability, current research explores leveraging Large Language Models (LLMs) to produce coherent textual explanations grounded in visual data, enabling interpretable decision-making across domains like healthcare, finance, and law [26]. These systems aim to translate convolutional layer activations into sub-symbolic reasoning steps that can be articulated in natural language. Huang et al. [27] introduced a physics-aware Deep Neural Network (DNN) model that improved transparency but sacrificed classification performance due to alterations in the physical properties of input images. Shajalal, Boden and Stevens [28] developed an XAI model focused on helping experts understand neural decisions through feature attribution, enhancing the interpretability of medical imaging applications. Similarly, Dasari and Bhukya [29] proposed a hybrid CNN-LSTM model (EdeepVPP) to automate feature extraction with built-in interpretability. However, the method's reduced complexity limited its performance on standard convolution operations. Luo et al. [30] presented a CNN-based Selective Fixed-Filter Active Noise Control (SFANC)

framework for real-time applications. While effective in noise reduction, it suffered from convergence issues and tracking limitations. B e g u m et al. [31] introduced a Lightweight CNN (LCNN) enhanced by LSTM layers to improve software defect detection. Although effective in prediction tasks, the model's applicability to image recognition was constrained by its narrow domain adaptation. The existing body of work underscores the need for a unified, explainable, and high-performance CNN framework. Addressing these gaps, our research proposes the DeConvolve explainability methodology, which leverages multimodal LLMs to extract sub-step explanations from CNN feature maps. This method aims to balance performance with interpretability, enhancing the transparency, trustworthiness, and practical deployment of CNNs in critical decision-making applications.

3. Proposed methodology

In this section, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to process Human-Cognizable Convolutional Networks and make decisions by attributing essential features. This enhances the cognitive alignment between ML outputs and human reasoning. The integration of LLMs helps extract human-understandable sub-steps and provide textual explanations for model inference.

3.1. Enforcing human cognizable sub-steps by analyzing the feature map

CNNs consist of convolutional layers that learn feature extraction from the training data [32]. Each filter acts as a selector, simplifying the sub-steps from a cognoscibility perspective and reducing the cognitive load on the LLM during post-analysis. The following modifications to the CNN training loss function are proposed to mitigate these issues: Firstly, image thresholding is performed on the feature map using a low threshold value to nullify pixels with extremely low focus (as measured by the heat map obtained from CAM). Then, the ratio of non-black pixels to the total image size is added to reduce the area of focus of the filter. This encourages filters to perform selection along the height and width axes, while the selection on the depth axis is split across the filters in the layer. Lastly, the dispersion of the areas of focus is incorporated into the cost function. This encourages each filter to observe a single spot along the image across all channels, as expressed by the next equation:

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \times \left(\frac{\text{Area}(\text{non}_{\text{black}})}{\text{Image}_{\text{size}}}\right) + \beta \times \sigma_{\text{non}_{\text{black}}}. \quad (1)$$

The next phase involves choosing the most pertinent features from their receptive fields in the input feature maps. The feature extraction manifests in two ways: transforming the input feature map into a more meaningful representation and focusing on the image regions through localization. While these transformations do not hinder explainability due to their visually apparent effects, localization complicates the inference process. Localization manifests as a selection along three axes of the image: height, width, and depth (channels, usually RGB in the input image). Localization along the depth axis cannot be visualized and hence, cannot be explained. However, localization along the height and width axes is crucial for

explainability, as it directs attention to visually and cognitively recognizable areas of the image. Extensive research has been conducted [10] to highlight the areas of interest in an image using the SHapley Additive exPlanations (SHAP) and GradCAM techniques. However, existing work on the explainability methodology provides an overview of model behaviour at the level, but does not delve deeper into the functioning of individual filters or neurons. The preliminary goal here is to obtain human-cognizable image patches of every filter. The CAM works by composing feature maps of various filters based on their impacts on the predictions. The feature map of each filter presents two main issues that hinder human cognizability: scattered regions of interest, which produce meaningless patterns across the image, and large areas of interest, where the filter is either redundant (performing no localization) or localizes along the depth axis. Moreover, the learned patterns from different filters can be conflicting. For instance, in a CNN used to identify human figures, if the first filter learns that the presence of a torso and head is positively linked to the prediction, and another filter learns that the presence of a torso is negatively linked to the output prediction, this conflict can be better represented by one filter focusing solely on the head, while the other filter remains blank. The terms are defined as follows:

- n is the number of samples in the dataset.
- y_i is the actual target value for sample i.
- $\hat{y}_i$ is the predicted target value for sample i.
- The non-black area of focus after applying the black threshold.
- Image size is the size of the image.
- σ_{non} is the explainability hyperparameter that controls the trade-off between the regular MSE loss and the additional terms.
- β is the scattering hyperparameter that controls the trade-off between the size of the receptive area and the dispersion of areas within the receptive field.

The modifications to the cost function can be visualized in Fig. 1. As the two new terms are progressively added, the feature map becomes more meaningful. A metric has been developed to quantify the concise learning described earlier. For each heatmap generated by a specific convolutional layer, the ratio of pixels with the highest entropy in the final classification to those with minimal impact on the prediction is calculated. A detailed evaluation of this metric is presented in the results section.

3.2. Scoring the filters

The process begins by scoring each filter, where the score represents the filter's influence on the predicted class. Specifically, it measures the correlation between the mean of pixels in the feature map and the output probability of the predicted class. To compute this score, a sub-model selection is performed. Fig. 1 represents the image with the area of focus highlighted with explainability terms, while Fig. 2 illustrates the flow of neurons in a human-cognizable manner. Fig. 3 shows how the pertinent filter convolution block is measured, while Fig. 4 demonstrates the extraction of a model with layers following the selected layer.

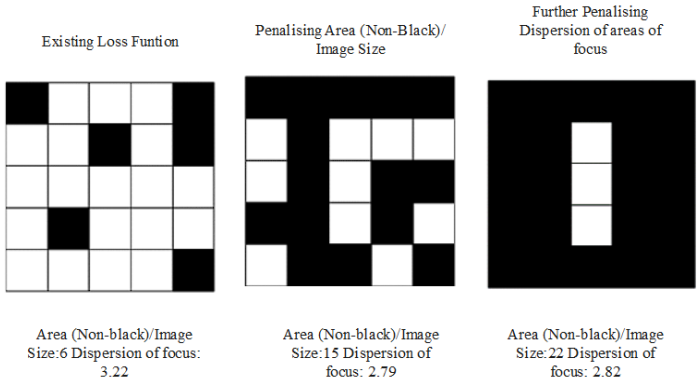


Fig. 1. A sample representation of images with ROI highlighted in white, with the explainability terms

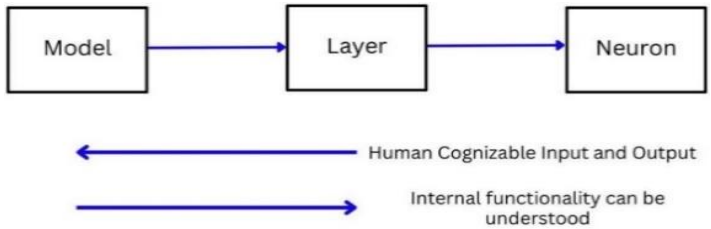


Fig. 2. Model's human-cognizable neuron flow

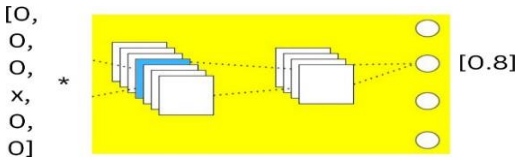


Fig. 3. Measurement of a filter convolution block

Measuring the change induced by constant matrices on the pertinent filter convolution block, with the matrices corresponding to all other filters in the block set to a null matrix. For the matrix associated with the relevant filter, a constant matrix is applied, beginning with a zero matrix and gradually increasing the constant until it reaches one. The predicted probabilities are then recorded, and the increase in probability is assigned as the score for the filter.

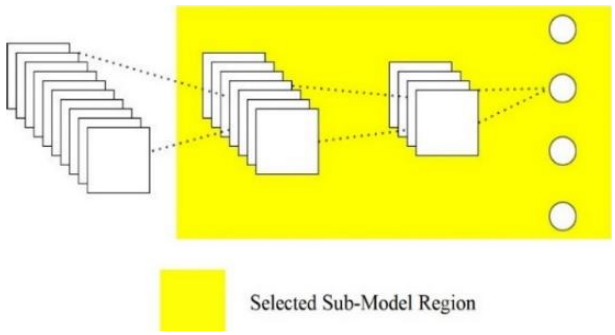


Fig. 4. Extraction of a model with layers after the selected layer

3.3. Obtaining activation maps

Inspired by class activation maps [8], the primary goal here is to generate heatmaps corresponding to each filter. This is achieved by performing gradient ascent on the inputs to the filter, using the corresponding class activation as the objective function. The heatmaps obtained are image-agnostic, evaluated, and stored as an extension of the model's training process.

3.4. Post hoc inference analysis using LLMs

We now analyse the network upon completion of inference, i.e., post-hoc explainability. The trained CNN model generates predictions for input images. In this step, we aim to convert the CNN's reasoning process into textual form. The contribution of each filter to the final decision is computed by simulating its activation and observing the change in output probability. This produces an influence score for each filter. These scores guide a language model in generating contextually relevant sentences that form a human-understandable explanation of the model's inference.

Algorithm 1. Scoring the impact of filters on the prediction

Step 1. num_layers ← : total number of layers in the CNN

Step 2. num_filters[num_layers] ← list storing the number of filters in each layer

Step 3. Score ← a matrix of dimensions num_layers × max(num_filters)

Step 4. For each layer in num_layers:

a. Define sub_model as the model sliced from the layer to the final output

b. For each filter in num_filters[layer]:

i. one_hot_input ← : a tensor of all zeros with dimensions equal to the input size of sub_model.get_layer[0]

ii. baseline_prob ← output probability from sub_model with one_hot_input

iii. Set one_hot_input[filter] ← unit (1-valued) tensor at the filter index

iv. activated_prob ← output probability from sub_model with modified one_hot_input

v. Score[layer][filter] ← activated_prob - baseline_prob

Step 5. Return the complete Score matrix

3.4.1. Sentential descriptions of individual feature maps

The transformation of image data into text presents significant challenges in computational tasks. Two main obstacles hinder the conversion process: firstly, identifying specific regions of interest within the three-dimensional image, and secondly, ensuring the conversion aligns effectively with the intended output requirements downstream. To pinpoint the areas of interest within the image, the previously stored heatmap is superimposed onto the input image. This superimposition creates a trade-off between contextual information and focus. Enhancing the clarity of the heatmap grid allows for greater focus on the relevant

object of interest. However, this sharpening process inadvertently reduces the model's awareness of its spatial surroundings and its role within the larger context. The extent to which sharpening is applied depends on the specific model used and the nature of the task being performed. Additionally, as shown in Fig. 5, two heatmaps are identified that convey the regions of focus unambiguously: The proportionate heatmap, where the intensity of every pixel is multiplied by the heatmap value, resulting in a pixel's opacity being proportional to its importance, and the encircled threshold heatmap, where the heatmaps are smoothed, multiplied with the input image, and all pixel values above a threshold are placed within a boundary. This boundary is drawn onto the image, allowing for a visually obvious compensation for the loss of contrast due to focus. The superimposed heatmaps are passed into a multimodal large language model with an appropriate prompt for context-aware labeling of the object.

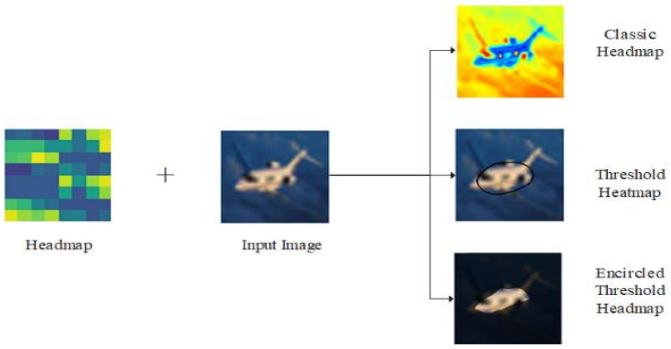


Fig. 5. Representation of ROIs

3.4.2. Semantic averaging of the sentences

The obtained textual explanations, in conjunction with the neuron scores retrieved (as computed in Section 3.2), are used to aggregate the sentential descriptions. This aggregation is limited by the intelligence that the current LLMs possess. This step involves the LLM understanding the interplay between various sentences as guided by the scores. At this stage, the LLM model presents a simple aggregation strategy, averaging sentences grouped by their scores. Then, the various bucket average sentences are averaged based on the mean score of the filter within the bucket. The aggregation strategies greatly affect the efficacy of DeConvolve, where specific tasks may require custom aggregation based on the logical "depth" of the task and the LLM's understanding of the topic.

4. Experiments and results

The proposed DeConvolve methodology demonstrates considerable promise in improving the interpretability of Convolutional Neural Networks (CNNs). The experiment is set up in a Python environment, Version 3.8 software tool, 16 GB RAM, Intel i7 Processor, Windows 10 operating system, with 16GB GPU and 1TB

SSD. To ensure the effectiveness of the custom loss function, a Convolutional Neural Network (CNN) model with an architecture containing 3 Convolutional Layers, 2 Max Pooling Layers is employed along a Global Average Pooling Layer is employed.

4.1. Improved feature maps on CIFAR-10

Table 1. Represented the feature mpat based on LLM

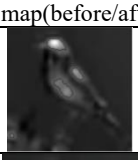
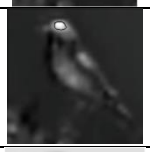
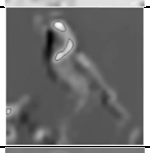
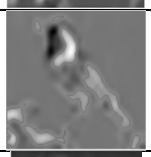
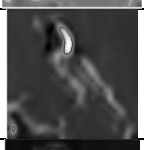
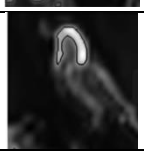
Filter	Ft. map(before/after)	LLM Description
1		The image shows a filter map highlighting the salient features of a bird, with the whitish area indicating the region of interest. The highlighted regions correspond to the bird's head and beak, which are the most significant parts
		The image patch primarily highlights a bird-like structure, with the most salient or "hot" area being around the head of the bird. The heatmap indicates that the CNN model is focusing on the upper portion of the bird, possibly identifying features such as the beak or the shape of the head
2		The hotter areas of the heatmap, particularly around the head and body, indicate that these regions are significant for the CNN model's prediction. The whitish areas suggest that these parts of the bird, including its general shape and position, are the primary focus influencing the output
		The filter map highlights a region resembling a humanoid figure with a pronounced facial structure, particularly the eyes and a smile. The hotter areas emphasize these features, suggesting the CNN is focusing on a part of the image that may include a face or face-like pattern, which could influence the classification decision
3		The filter map highlights a region of interest, predominantly a whitish area, which seems to outline the shape of a bird. The salient features include a defined head and beak area, with the contrast in the heatmap indicating important structural elements like the body and possibly wings
		The filter map highlights a whitish area prominently in the upper central region, which appears to outline the curved shape of the bird's head or body
4		The filter map highlights a region with a distinct whitish hue, indicating that it is the primary area of interest. This region appears to outline a bird's head and beak, which is crucial for the CNN model's predictions. The key features captured here are likely to help identify this image as a bird
		The image depicts a bird, with the highlighted whitish area indicating the region of interest. The hotter area in the heatmap, marked by a distinct U-shaped structure, is the bird's head, suggesting that this feature is crucial for the CNN model's prediction

Table 1 shows four feature maps highlighting the improvements in human cognizable representations. The filter maps are shown as before and after the application of the custom loss function. Additionally, the immediate increase in the relevance of the generated text is observed. The image patches that are most salient

and likely to influence the output are described, with a key focus on the hotter areas highlighted in the heat map, as represented in Fig. 5.

The CIFAR-10 dataset consists of 60,000 32×32 color images divided into 10 classes, with each class containing 6000 images. The filter sizes used in the convolutional layers were 32 filters of size 3×3, 64 filters of size 5×5, and 128 filters of size 3×3. The hyperparameters used are:

- Learning Rate: 0.001.
- Batch Size: 64.
- Epochs: 50.
- Optimizer: Adam.
- Loss Metric: Sparse Categorical Cross-entropy.
- Explainability Hyperparameter (α): 0.1.
- Scattering Hyperparameter (β): 0.05.

4.2. Improvement in succinct learning

The suggested methodology refines and simplifies learnt patterns in CNN feature maps to enhance interpretability through Succinct Learning. This process involves thresholding feature maps to create binary heatmaps, aggregating them to identify positions with consistently high activations, and processing these positions to accumulate weights associated with significant activations. The resulting processed matrix highlights key patterns, eliminating less relevant information and minimizing conflicts between filters. For example, in identifying human figures, one filter might learn that the presence of a torso and head is positively linked with the prediction, while another might associate the torso negatively. By ensuring each filter focuses on distinct patterns, conflicts are reduced, and the cognitive load during post hoc analysis is alleviated on the CIFAR-10 dataset. Furthermore, an improvement in the clarity of learned patterns is observed where the ratio of significantly fewer activations is increased from 0.85 before applying the heuristic to 1.00. This demonstrates the heuristic's effectiveness in achieving a succinct and interpretable representation. Please refer to the supplementary material for the algorithm used to calculate the ratio.

4.3. Example explanations

This study utilizes the CIFAR-10 dataset to test the explainability of the CNN by generating textual explanations through an LLM. Specifically, the feature maps from the CNN are fed into the LLM to generate detailed explanations for image classifications. The study focuses on two classes from CIFAR-10: automobiles and airplanes. The results for images from both these classes are discussed below.

4.3.1. Misclassification of car image

When a car image, as shown in Fig. 1, is input into the CNN, it is incorrectly classified. The generated textual explanation (Place-holder for Explanation 1) mentions the absence of a front bumper, which is highlighted as the key reason for misclassification. However, upon examining the image, it is evident that the front

bumper is present, indicating a significant discrepancy between the actual image features and the explanation provided by the model.

As shown in Fig. 6, by combining positive and negative filter maps, an understanding of how the model identifies a vehicle is gained. When there is an identification, the model relies heavily on clear and distinct features such as the hood, front bumper, roofline, wind-downs, trunk, and rear bumper to classify the vehicle. On the other hand, when there is a negative identification, the model struggles with less-defined features, such as the bumper and grille, wheels, and diffuse shapes of the roof and windows. These areas negatively impact the model's ability to correctly identify the vehicle. Here, the most important substructures present are, Front End (Hood and Front Bumper), Cabin Area (Roof and Windows), Rear End (Trunk and Rear Bumper), while the most important substructures absent include Front/Rear End (Bumper and Grille), Wheels, Cabin Area (Roof and Windows), which are diffuse and less distinct. The identification process is significantly influenced by the presence of clear and distinct vehicle features. The importance of well-defined structural details is observed carefully to ensure explainability. To investigate further, another car image with a prominently visible front bumper, like in Fig. 7, is passed through. This time, the CNN correctly classifies the image as a car. The explanation provided by the LLM is correctly identified by identifying the front bumper as a distinguishing feature, which aligns with the visual attributes of the image. The classification of airplane images in Fig. 8 is performed based on the filter maps with positive correlation features to the output class prediction, where the model shows higher activation in areas that outline specific structures. In this case, the most activated region resembles the hull of a boat. The elongated shape and distinct outline are key features that the model identifies, enabling it to predict the output class with higher confidence.

Fig. 6. Incorrectly classified automobile

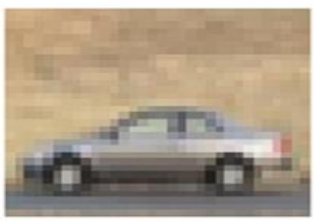


Fig. 7. Correctly classified automobile

Fig. 8. Correctly classified airplane

The hull-like structure of a boat is a salient feature that significantly influences the model's decision. For the filter maps with a negative correlation to the output class prediction, the model shows higher activation in areas that seem to detract from the confidence in the predicted class. The most prominent feature in these activated areas is the outline of the animal, which suggests that it is most likely a horse. The model appears to be agnostic to areas that do not contain significant or distinct features, resembling either the hull of a boat or the outline of a horse. These regions do not activate strongly in either positive or negative filter maps, indicating that they do not contribute significantly to the model's decision-making process. Key features in a positive explanation include the hull-like structure of a boat, elongated shape, and the distinct outline. The absence of a negative explanation affects the outline of the horse, legs, and body shape. In summary, the model highlights the hull-like structure of a boat as a crucial feature for predicting the output class. Conversely, the presence of horse-like features negatively impacts prediction. The model's decision-making process is heavily influenced by these salient structures, while areas without such distinct features do not significantly affect the output. Large Language Models (LLMs) are then employed to produce textual descriptions of these regions, resulting in a cohesive explanation of the decision-making process of the model and improvement in succinct learning, as represented in Figs 6-9.

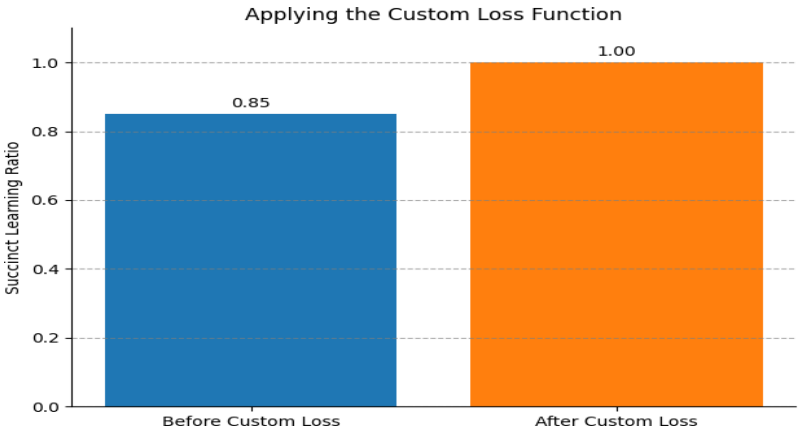


Fig. 9. Improved succinct learning

4.4. Discussion

The proposed DeConvolve methodology demonstrates a superior performance with improved interpretability of the CNNs. The algorithm aims to effectively process and analyze feature maps generated by CNNs. This algorithm identifies and quantifies significant activation patterns within these feature maps, facilitating easier interpretation and post hoc analysis of the CNN's learned representations. The process begins by thresholding each feature map from a CNN layer to create binary heat maps. These binary heatmaps indicate whether each element in the feature map exceeds a specified activation threshold, thus highlighting areas of significant activation. The next phase involves identifying high-count positions in the combined matrix. The final output of the algorithm is the calculated ratio, which indicates the proportion of significant activations relative to the less significant ones. There are enhancements made to the training process of the CNN by integrating feature map abstraction at the filter level for enhanced human recognizability. This is accomplished by neutralizing low-focus pixels and modifying the loss function to prioritize smaller, more concentrated regions of focus. The effectiveness of the proposed approach in presenting extensive and sophisticated rationales underlying CNN inferences is proven through trials on the CIFAR-10 dataset for validating these developments. In addition, to enhance the model's ease of use and ability to replicate results, a web interface that is intuitive and accessible to users is created. This interface facilitates convenient testing and display of the model's explanations, providing a wide range of modification possibilities to enhance interpretation. The findings emphasize the need to use DeConvolve to enhance the transparency and comprehensibility of CNNs, thereby facilitating the model's responsible implementation in diverse applications.

5. Conclusion

The proposed DeConvolve methodology showcases a commendable performance with improved interpretability of CNNs. LLMs are then employed to produce textual descriptions of these regions, resulting in a cohesive explanation of the model's decision-making process. This combination ensures that the feature maps undergo careful pre-processing at the model and image processing level. The heat maps for each filter are generated using GradCAM, and these heat maps are optimally represented for LLM comprehension. The Grad-CAM aids in making the AI system compliant with regulations of the General Data Protection Regulation (GDPR) and AI transparency, making it adaptable to text data using mapping word or sentence embeddings to the corresponding heatmaps. Finally, the explanation for the specific images is derived through weighted semantic averaging for data processing by image format, class, and rationale. This enables professional classification using CLIP training for generating robust explorations. In future work, the power of CLIP models can be harnessed to improve learning explanations in a supervised model, which is currently impossible due to the unavailability of datasets with images categorized based on their classes and explanations. For example, X-ray scans of the lungs are used to detect the presence of a cancerous

tumor, with an oncologist providing an analysis of the observations using X-ray images. Furthermore, Language models used here can be pre-trained on lung cancer and its visual characteristics. Fig. 1 depicts a possible architecture. The CLIP model was proposed by Radford and other authors (2021) in the work titled "Learning Transferable Visual Models From Natural Language Supervision". The proposed loss function works as an antecedent to training explainability models upon the creation of datasets.

References

1. Huu, H. D., N. M. Bui, V. P. Hoang, T. B. Quy, Y. H. Thi. Improved Convolutional Neural Network-Based Bearing Fault Diagnosis Using Multi-Phase Motor Current Signals. – International Journal of Electrical & Computer Engineering, Vol. **15**, 2025, No 2, pp. 1656-1669.
2. Wijaya, R., G. Kosala. Stress Detection through Wearable Sensors: A Convolutional Neural Network-Based Approach Using Heart Rate and Step Data. – International Journal of Electrical & Computer Engineering, Vol. **15**, 2025, No 2, pp. 1880-1888.
3. Vaquerizo-Villar, F., G. C. Gutiérrez-Tobal, E. Calvo, D. Álvarez, L. Kheirandish-Gozal, F. Del Campo, D. Gozal, R. Hornero. An Explainable Deep-Learning Model to Stage Sleep States in Children and Propose Novel EEG-Related Patterns in Sleep Apnea. – Computers in Biology and Medicine, Vol. **165**, 2023, 107419.
4. Ren, Z., K. Qian, F. Dong, Z. Dai, W. Nejdl, Y. Yamamoto, B. W. Schuller. Deep Attention-Based Neural Networks for Explainable Heart Sound Classification. – Machine Learning with Applications, Vol. **9**, 2022, 100322.
5. Mallampati, S. B., H. Seetha. Enhancing Intrusion Detection with Explainable AI: A Transparent Approach to Network Security. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 1, pp. 98-117.
6. Camacho, M., M. Wilms, P. Mouches, H. Almgren, R. Souza, R. Camicioli, Z. Ismail, O. Monchi, N. D. Forkert. Explainable Classification of Parkinson's Disease Using Deep Learning Trained on a Large Multi-Center Database of T1-Weighted MRI Datasets. – NeuroImage: Clinical, Vol. **38**, 2023, 103405.
7. Altan, G. DeepOCT: An Explainable Deep Learning Architecture to Analyze Macular Edema On OCT Images. – International Journal of Engineering Science and Technology, Vol. **34**, 2022, 101091.
8. Sun, C., H. Xu, Y. Chen, D. Zhang. AS-XAI: Self-Supervised Automatic Semantic Interpretation for CNN. – Advanced Intelligent Systems, Vol. **6**, 2023, No 12, 2400359.
9. Sudha, V. K., D. Kumar. A Hybrid CNN and LSTM Network for Heart Disease Prediction. – SN Computer Science, Vol. **4**, 2023, No 2, p. 172.
10. Górriz, J. M., I. Álvarez-Illán, A. Álvarez-Marquina, J. E. Arco, M. Atzmueller, F. Ballarini, E. Barakova, G. Bologna, P. Bonomini, G. Castellanos-Dominguez, D. Castillo-Barnes. Computational Approaches to Explainable Artificial Intelligence: Advances in Theory, Applications and Trends. – Information Fusion, Vol. **100**, 2023, 101945.
11. Yang, D., H. R. Karimi, L. Gelman. An Explainable Intelligence Fault Diagnosis Framework for Rotating Machinery. – Neurocomputing, Vol. **541**, 2023, 126257.
12. Bhandari, M., P. Yogarajah, M. S. Kavitha, J. Condell. Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP. – Applied Sciences, Vol. **13**, 2023, No 5, p. 3125.
13. Naeem, H., B. M. Alshammari, F. Ullah. Explainable Artificial Intelligence-Based IoT Device Malware Detection Mechanism Using Image Visualization and Fine-Tuned CNN-Based Transfer Learning Model. – Computational Intelligence and Neuroscience, Vol. **2022**, 2022, No 1, 7671967.

14. Sanakkayala, D. C., V. Varadarajan, N. Kumar, Karan, G. Soni, P. Kamat, S. Kumar, S. Patil, K. Kotecha. Explainable AI for Bearing Fault Prognosis Using Deep Learning Techniques. – Micromachines, Vol. **13**, 2022, No 9, 1471.
15. Pradhan, B., R. Jena, D. Talukdar, M. Mohanty, B. K. Sahu, A. K. Raul, K. N. Abdul Maulud. A New Method to Evaluate Gold Mineralisation-Potential Mapping Using Deep Learning and an Explainable Artificial Intelligence (XAI) Model. – Remote Sensing, Vol. **14**, 2022, No 18, p. 4486.
16. Nazari, M., A. Kluge, I. Apostolova, S. Klutmann, S. Kimiaei, M. Schroeder, R. Buchert. Explainable AI to Improve Acceptance of Convolutional Neural Networks for Automatic Classification of Dopamine Transporter SPECT in the Diagnosis of Clinically Uncertain Parkinsonian Syndromes. – European Journal of Nuclear Medicine and Molecular Imaging, Vol. **49**, 2022, pp. 1176-1186.
17. Shelke, N., S. Maurya, R. Ithape, Z. Shaikh, R. Somkunwar, A. Pimpalkar. Towards an Automated Weather Forecasting and Classification Using Deep Learning, a Fully Convolutional Network, and Long Short-Term Memory. – International Journal of Electrical and Computer Engineering, Vol. **15**, 2025, No 2, pp. 1868-1879.
18. Ghrabat, M. J. J., Z. A. Hussien, M. S. Khalefa, Z. A. Abduljabbar, V. O. Nyangaresi, M. A. Al Sibahee, E. W. Abood. Fully Automated Model on Breast Cancer Classification Using Deep Learning Classifiers. – Indonesian Journal of Electrical Engineering and Computer Science, Vol. **28**, 2022, No 1, pp. 183-191.
19. Nafisah, S. I., G. Muhammad. Tuberculosis Detection in Chest Radiograph Using a Convolutional Neural Network Architecture and Explainable Artificial Intelligence. – Neural Computing and Applications, Vol. **36**, 2024, No 1, pp. 111-131.
20. Hossain, M. M., M. S. Ali, M. M. Ahmed, M. R. H. Rakib, M. A. Kona, S. Afrin, M. K. Islam, M. M. Ahsan, S. M. R. H. Raj, M. H. Rahman. Cardiovascular Disease Identification Using a Hybrid CNN-LSTM Model with Explainable AI. – Informatics in Medicine Unlocked, Vol. **42**, 2023, 101370.
21. Gunashekar, D. D., L. Bielak, L. Hägele, B. Oerther, M. Benndorf, A. L. Grosu, T. Brox, C. Zamboglou, M. Bock. Explainable AI for CNN-Based Prostate Tumor Segmentation in Multi-Parametric MRI Correlated to Whole Mount Histopathology. – Radiation Oncology, Vol. **17**, 2022, No 1, p. 65.
22. Fanizzi, A., M. C. Comes, S. Bove, E. Cavalera, P. de Franco, A. di Rito, A. Errico, M. Lioce, F. Pati, M. Portaluri, C. Saponaro. Explainable Prediction Model for the Human Papillomavirus Status in Patients with Oropharyngeal Squamous Cell Carcinoma Using CNN on CT Images. – Scientific Reports, Vol. **14**, 2024, No 1, 14276.
23. Sarkar, O., M. R. Islam, M. K. Syfullah, M. T. Islam, M. F. Ahamed, M. Ahsan, J. Haider. Multi-Scale CNN: An Explainable AI-Integrated Unique Deep Learning Framework for Lung-Affected Disease Classification. – Technologies, Vol. **11**, 2023, No 5, p. 134.
24. Bao, J., M. Ye. Head Pose Estimation Based on Robust Convolutional Neural Network. – Cybernetics and Information Technologies, Vol. **16**, 2016, No 6, pp. 133-145.
25. Railkar, D., S. Joshi. AHT-QCN: Adaptive Hunt Tuner Algorithm Optimized Q-learning Based Deep Convolutional Neural Network for the Penetration Testing. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 3, pp. 182-196.
26. Pang, P., J. Tang, J. Luo, M. Chen, H. Yuan, L. Jiang. An Explainable and Lightweight Improved 1D CNN Model for Vibration Signals of Rotating Machinery. – IEEE Sensors Journal, Vol. **24**, 2024, No 5, pp. 6976-6997.
27. Huang, Z., X. Yao, Y. Liu, C. O. Dumitru, M. Datcu, J. Han. Physically Explainable CNN for SAR Image Classification. – ISPRS Journal of Photogrammetry and Remote Sensing, Vol. **190**, 2022, pp. 25-37.
28. Shajalal, M., A. Boden, G. Stevens. Explainable Product Backorder Prediction Exploiting CNN: Introducing Explainable Models in Businesses. – Electronic Markets, Vol. **32**, 2022, No 4, pp. 2107-2122.
29. Dasari, C. M., R. Bhukya. Explainable Deep Neural Networks for Novel Viral Genome Prediction. – Applied Intelligence, Vol. **52**, 2022, No 3, pp. 3002-3017.

30. L u o, Z., D. S h i, J. J i, X. S h e n, W. S. G a n. Real-Time Implementation and Explainable AI Analysis of Delayless CNN-Based Selective Fixed-Filter Active Noise Control. – Mechanical Systems and Signal Processing, Vol. **214**, 2024, p. 111364.
31. B e g u m, M., M. H. S h u v o, M. K. N a s i r, A. H o s s a i n, M. J. H o s s a i n, I. A s h r a f, J. U d d i n, M. A. S a m a d. LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI. – IEEE Access, Vol. **12**, 2024, pp. 55744-55756.
32. H a m i d j a, K. B., F. W. R. T o k p a, V. M o n s a n, S. O u m t a n a g a. A Constrained Convolutional Neural Network with an Attention Mechanism for Image Manipulation Detection. – International Journal of Electrical & Computer Engineering, Vol. **15**, 2025, No 2, p. 2304.

Received: 14.04.2025. Revised version: 16.05.2025. Second Revision 26.06.2025. Accepted: 30.06.2025