Explaining Convolutional Neural Networks via a Concise and Hierarchical Approach

Jiawei Pan[1,2], Xiangyu Zhu[1,2], Haoyuan Zhang[1,2], Stan Z. Li[3], Zhen Lei[1,2,4,5,*]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Westlake University, Hangzhou, China
[4]CAIR, HKSIS, Chinese Academy of Sciences, Hong Kong, China
[5]SCSE, the Faculty of Innovation Engineering, M.U.S.T, Macau, China
{panjiawei2023, xiangyu.zhu, zhanghaoyuan2023, zhen.lei}@ia.ac.cn
stan.zq.li@westlake.edu.cn

Abstract

Explainable artificial intelligence (XAI) aims to bring transparency to black-box neural networks. Many innovative explainable methods provide rich and multifaceted explanations. However, these methods often have complex mechanisms or a high complexity of explanations, making them difficult for people to understand. To address this challenge, we introduce a concept-based explainable method which can improve the readability of deep neural network explanations by obtaining a concise set of high-quality explanations. We reduce the explanation redundancy by weighted hierarchical clustering, thereby obtaining a set of explanations that completely describe the input image and are crucial to the network's decision-making. Compared with existing XAI methods, our approach reduces the complexity while ensuring the integrity and comprehensibility of the explanation. In addition, we show that our approach can be traced back to the neuron-level explanation, which can also provide inspiration for model researchers to interpret the model. We validated the effectiveness of our approach through experiments in bird classification and facial recognition tasks. Specifically, we employed XAI to investigate the mechanisms behind face recognition, identifying critical neurons that correspond to key concepts in the recognition process.

1. Introduction

The increasing research interest in Explainable Artificial Intelligence (XAI) is driven by the pressing need for robust and reliable neural networks. As deep learning systems are progressively utilized in safety-critical areas like autonomous driving and biomedical diagnostics, the requirement for model transparency has soared to unprecedented heights. In order to meet this ever-growing demand, a series of explainable methods have been proposed. However, in explaining the decision-making process, the conciseness and comprehensibility of the explanation itself become a new challenge.

The post hoc XAI [29] method focuses on explaining the network decision-making process and the results of the decisions after model training. There are two main types of solutions to this problem: local methods explain a single input image in the form of an attribution map, thus determining 'where' important features appear (*e.g.* [23], [27], [17], [22], [11]), and global methods such as [2], [9], [10], [14] can visualize the concepts learned by the model and illustrate 'what' features the model is paying attention to. Recently, the work of [1], [15], [32], [6], [21] *et al.* has sought to integrate local and global methods. Among them, CRP [1] treats each neuron as an independent semantic feature detector and introduces the concept of a conditional mask relevance flow to generate attribution maps of specific neurons.

As explainable methods are constantly being updated, the complexity of the explanations themselves has gradually become an increasing concern. Taking the CRP [1] method as an example, although the CRP method completely shows which concept the model focuses on when making decisions, the explanations it produces are very complex. For example, taking Resnet-34 [12] as an example, CRP generates an explanation attribution map for each neuron in each convolutional layer, which means 8512 attribution maps. Although this exhaustive analysis retains all features influencing the decision, the resulting cognitive load makes it

*Corresponding author

difficult for amateur users to comprehend, thus fundamentally undermining the purpose of XAI to provide actionable insights.

To address this critical issue, we propose a novel concept-based explainable method that demonstrates key features discovered by the model during the decision-making process through a complete yet concise set of explanations. Our goal is to provide individuals with a set of intuitive and readable explanations. To this end, we reduce the redundancy caused by the encoding of similar features by network neurons during the explanation generation process, while ensuring the completeness of the explanations. As illustrated in Figure 1, our approach aggregates the original 8,512 redundant explanations produced by CRP [1] on ResNet-34 into approximately 25 representative semantic concepts. Although a related work [15] reduces complexity by incorporating non-negative matrix factorization (NMF) through cross-layer aggregation, its simple clustering approach loses the crucial connection between concepts and neurons. Specifically, concepts after dimensionality reduction no longer map directly to specific neurons - a core aspect of XAI that makes this simplification suboptimal. Our method is based on the conditional attribution map generated by CRP [1]. As illustrated in Figure 2(a), weighted hierarchical clustering is used to reduce the redundancy of explanations, and a structural similarity weight is introduced to make the clustering results more closely aligned with the original model structure, reducing the complexity of the explanations while improving their completeness. In addition, our method keeps the connection between concepts and neurons after dimension reduction; that is to say, each concept explanation can be traced back to specific neurons. As illustrated in Figure 2(b), each final concept (e.g., the "bright orange bill" concept) obtains a link with a group of neurons in the original deep network. Therefore, these traceable explanations serve as an effective network summary, facilitating further research on the model. We conducted quantitative and qualitative experiments in bird classification and face identity recognition tasks to prove the effectiveness of our method.

The main contributions of this work can be described as follows:

1) We propose a concept-based explainable method that reduces explanation redundancy while ensuring explanation completeness through weighted hierarchical clustering, obtaining a concise and complete set of explanations, thereby improving the readability of neural network explanations.

2) We verify that the concepts found by our approach are easy for people to intuitively comprehend and reduce redundancy while ensuring the completeness of the explanation.

3) We show that the proposed approach can provide inspiration for researchers to further interpret the network.

2. Related Works

XAI methods in general can be categorized into two types. One is the development and design of more transparent decision-making algorithms, such as [31], [30], [3], *et al.*, which develop and train novel models with intrinsic interpretability. The other type is to explain trained deep models post hoc, such as [23], [27], [17], [22], [11], [20], *et al.* In this paper, we focus on the second type, and we will discuss the existing post hoc XAI methods in terms of whether the explainable methods are based on Attribution or Concept.

2.1. Attribution-based methods

Attribution-based methods are commonly used as post-hoc explainable artificial intelligence (XAI) to determine the input variables that contribute to model predictions by generating saliency maps. These methods focus on providing explanations for the decision-making process of individual prediction instances and are also known as local methods. Local methods typically calculate the gradient sensitivity of input features or the response to input perturbations, effectively identifying local regions that play a critical role in a specific prediction process and revealing which input components have a significant impact on the model's decision.

One of the most classical algorithms is the CAM [33] method and its many modified versions.The core idea of CAM is to generate spatial attention heat maps related to the classification results by weighted fusion of the feature maps output from the last layer of convolution of a CNN. Grad-CAM [23] is a gradient-based method that computes the importance weights of each channel by using the gradient information of the target category scores with respect to the last convolutional layer's feature map. Grad-CAM++ [4] is an enhanced version of Grad-CAM, which improves the gradient-weighting mechanism to more accurately locate multiple critical regions and enhance the explainability of the heat map. Layer-CAM [13] is a class-activated mapping method based on multi-layer gradient information, which improves the localization accuracy by integrating shallow and deep convolutional features.XGrad-CAM [7] is a theoretically optimized gradient-weighted class-activated mapping method to generate the heat maps by improving the calculation of the gradient weights.

However, image gradients can only reflect how the model operates in an infinitely small neighborhood around the input, which can lead to misleading importance estimates [8]. In addition, local XAI methods also have an obvious problem: feature representation ambiguity. That is, although attribution graphs can reflect the correlation between input features and predictions, they cannot reveal the semantic-level features extracted by the model in key areas.

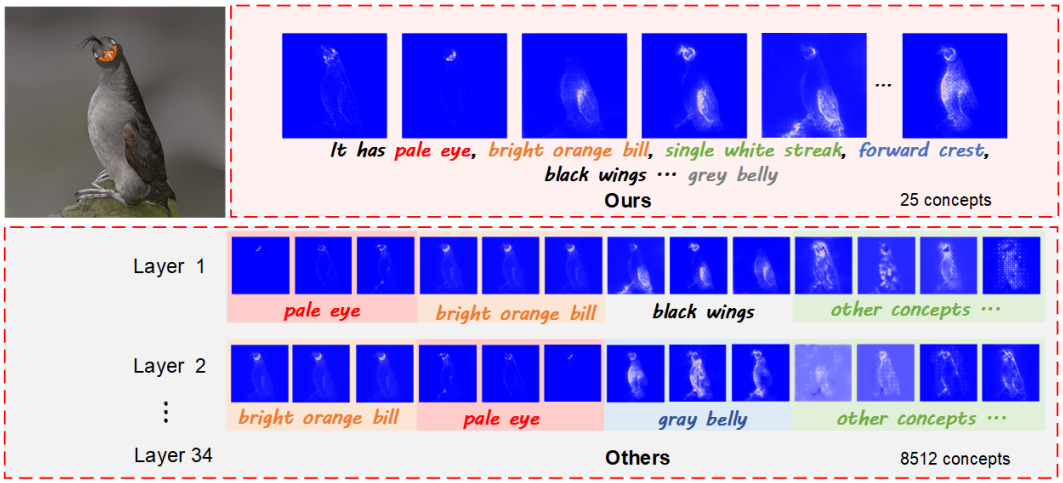

Figure 1. Many existing explainable methods generate explanations with high complexity, such as CRP [1] which produces 8512 explanations with duplicates. Our approach obtains approximately 25 representative high-level semantic concepts, thus significantly reducing complexity while ensuring the quality of the explanation.

2.2. Concept-based methods

Concept-based explainable methods, such as [2], [10], [14], aim to provide explanations beyond attribution-based methods by measuring the impact of preselected concepts on model output. Although this method appears to be more human-interpretable than standard attribution techniques, it requires a manually edited image database describing relevant concepts. [9] and other works have further extended this method to extract concepts without human supervision. Although such explanation methods can systematically reveal the decision-making basis of models and are also referred to as global methods, their limitation lies in the lack of instance-level explainability. For example, when faced with complex scenarios such as occlusion and lighting changes, global feature explanations struggle to dynamically reflect the decision-making path of specific samples. This shortcoming has prompted subsequent research to attempt to fuse the complementary advantages of global semantic features and local saliency analysis.

Recently, [1], [15], [32], [6], [21] *et al.* attempt to combine local and global methods. The integrated gradient method proposed by [21] quantifies the contribution of feature channels to prediction results, achieving the measurement of the importance of the concept at the instance level. The core of this method lies in the construction of concept activation vectors (CAV) and the modeling of predefined semantic concepts in a latent space using a linear classifier. However, its limitations include the need for a manually labeled concept image dataset as a supervisory signal and the lack of visualization capabilities in the original pixel space, making it difficult to intuitively present the spatial distribution patterns of concepts in input images. [1] proposes a framework for the detection of semantic features at the neuron level. This study treats the activation units of convolutional neural networks as independent semantic detectors and generates neuron-specific attribution maps through conditional relevance propagation (CRP). This method innovatively introduces a correlation score metric to extract consistent features of neuron activation patterns across multiple images, revealing the global semantic concepts encoded by neurons. However, the computational complexity of CRP is extremely high.

To reduce the complexity of the CRP method, recent studies have explored some approaches. [15] aims to reduce the dimensionality by aggregation across layers and non-negative matrix factorization (NNMF), and establish associations between the reduced-dimensional features and the input images by identifying attribution maps in the original database that are similar to the reduced-dimensional features. Although this method effectively accomplishes the task, it blurs the connection between the final concise interpretation and the specific neurons in the neural network. In other words, even if the feature maps generated after dimensionality reduction are similar to the attribution graphs produced by neurons, the reduced concepts no longer directly originate from specific neurons (and may be combinations or decompositions of concepts).

3. Method

This section begins with a review of the CRP [1] method and describes the specifics for adopting hierarchical clustering [18] and the connection between concepts and neurons.

3.1. CRP Review

Given an image x and a trained network $F(x)$, an explainable method $E(x)$ such as CRP [1] will construct a

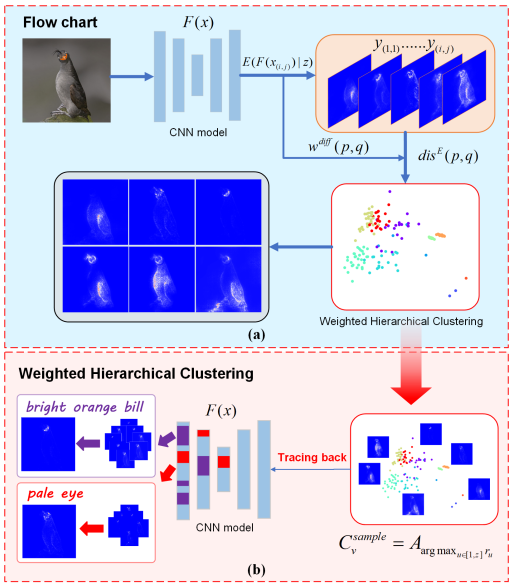

Figure 2. (a): Flow chart of our method. We first input the image into a trained network, generate attribution maps through concept-based explainable method, and then conduct the weighted hierarchical clustering algorithm to obtain a set of explanations. (b): The explanations generated by our method can be traced back to specific neurons.

conditional attribution map $y_{(i,j)}$ based on the classification category z for each neuron $x_{(i,j)}$ (neuron j in layer i) in the neural network, as shown below:

$$y_{(i,j)} = E(F(x_{(i,j)})|z), \tag{1}$$

where the CRP method generates n conditional attribution maps for n neurons in the neural network, which presents a challenge in systematically explaining the network, since even the simplest CNN can produce a large number of explanations. CRP [1] notes that the attribution maps at different layers of the network exhibit a flow of relevance, meaning that the concept explanations produced by neurons in the shallower layers can be combined with concepts in the deeper layers. This observation aligns with our existing understanding of neural networks. In other words, the n concept attribution maps generated by CRP inherently carry a hierarchical relationship. Therefore, we employ hierarchical clustering to cluster them and introduce weights to reflect the importance of each concept. To better understand this process, as shown in Figure 2(a), the image n is first input into the trained network $F(x)$, and attribution maps are generated through a concept-based explainable method E. Then, a weighted hierarchical clustering algorithm is conducted to obtain a set of high-quality explanations.

3.2. Weighted Hierarchical Clustering

We employ hierarchical clustering algorithms to process the concept attribution maps generated by the Concept Relevance Propagation (CRP) method. As illustrated in Figure 2(a), this algorithm utilizes a bottom-up agglomerative hierarchical clustering strategy, iteratively merging the most similar clusters to construct a hierarchical tree structure. This approach effectively reveals the intrinsic hierarchical relationships within the data. Specifically, for a concept attribution map generated by n neurons (where $n = 8,512$ in the ResNet-34 architecture), the algorithm initializes by treating each attribution map as an independent cluster, fully preserving its original spatial activation patterns and relevance scores. After initialization, the algorithm identifies pairs of similar clusters based on a specific distance metric and linkage criterion through iterative computation, eventually forming a complete hierarchical structure or achieving a preset termination condition. The clustering process can be described as:

$$dis^E(p,q) = \sqrt{(y_p - y_q)^2}, \tag{2}$$

$$w^{diff}(p,q) = |r_p - r_q|, \tag{3}$$

$$dis(p,q) = \lambda * dis^E_{norm}(p,q) + (1-\lambda) * w^{diff}_{norm}(p,q), \tag{4}$$

where p and q denote individual neurons, dis^E is the Euclidean distance, and w^{diff} refers to the difference weight. In CRP [1], when generating a concept explanation y_p for each neuron x_p, it will also generate a relevance r_p between the explanation and the prediction label. The higher this relevance, the greater the importance of this concept in model decision making. In addition, the deeper the neurons, the higher the likelihood of high relevance. Due to the hierarchical structure mentioned above, the higher the degree of similarity in relevance, the higher the likelihood that they are located in close proximity in the network hierarchy. Therefore, we add a weight w^{diff} to the Euclidean distance dis^E to represent structural similarity, and introduce a parameter λ to control the degree of influence of w^{diff} on the distance metric. In addition, the subscript $norm$ refers to normalization.

The distance calculation method can be arbitrarily selected and is usually determined by the specific distribution of the data set. Take the popular shortest distance method (single link method) as an example:

$$d_{min}(C_i, C_j) = min_{p\in C_i, q\in C_j} dis(p,q), \tag{5}$$

where $dis(p,q)$ is the weighted Euclidean distance mentioned above. It is important to note that when selecting certain distance calculation methods (such as the Ward method [25]), specific factors must be considered: the geometric consistency needs to be maintained. Our weighted distance

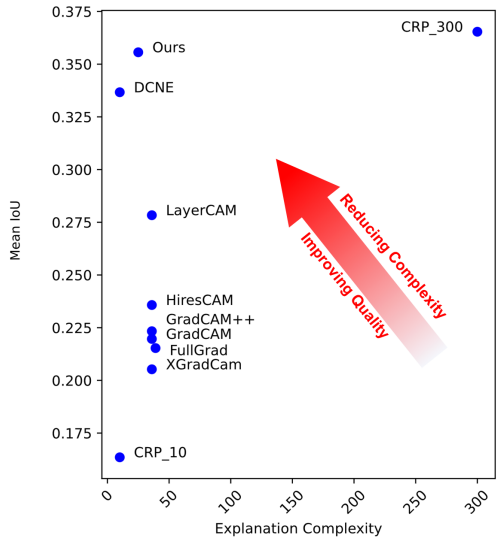

Figure 3. Comparison of mean IoU values and the explanation complexity of XAI methods.

metric is non-negative and satisfies the triangle inequality (since both dis^{E}_{norm} and w^{diff}_{norm} are non-negative and linearly combined). Therefore, our weighted Euclidean distance can be applied to the Ward method.

Finally, we use the silhouette coefficient [24] to evaluate the quality of the clustering and search for the largest silhouette coefficient within the preset clustering range as the stopping condition for hierarchical clustering to obtain the optimal number of clusters.

3.3. Connection between Concepts and Neurons

In the previous description, we mentioned that the CRP [1] method generates n concept attribution maps for n neurons conditional on the categorical class. After weighted hierarchical clustering, these attribution maps become the optimal clustering of m concepts $C_v, v \in [1, m]$. Suppose that there are z attribution maps $A_u, u \in [1, z]$ in the concept cluster C_v. We selected the attribution map with the highest relevance value in these attribution maps as a representative sample of this $concept C_v^{sample}$, which can be noted as:

$$C_v^{sample} = A_{\arg\max_{u \in [1,z]} r_u} \tag{6}$$

Thus, we set a link between the reduced-dimensionality concept and the single attribution map, and the single attribution map corresponds to a unique neuron. Therefore, our method can be used to trace the concept back to a specific neuron. If further network interpretation is required, we can obtain all neurons corresponding to the attribution maps of the concept C_v.

4. Experiment

In this section, we conduct quantitative and qualitative experiments to verify that our method can reduce the complexity of explainable methods while ensuring the quality of the explanation. In addition, we discuss how explanations that can be traced back to neurons provide the possibility for further manipulation of the network.

4.1. Datasets

Our explainable method, like most XAI methods, can be applied to any model based on convolutional networks. We take ResNet-34 [12] as an example and use the expert semantic annotation performed by [15] on the CUB-200-2011 (CUB) dataset [28] for our experiments. The original dataset (CUB) consists of images of 200 different bird species, while the new CUB-expert [15] dataset is expert annotated for several selected bird species, with 2-4 features annotated pixel-wise for each species and multiple masks created for each image to represent important features. In addition, we chose a face identity and attribute recognition model [19] for qualitative experiments, which is a ResNet-18 model trained on the CelebAMask-HQ dataset [16]. The face data used for identification contains 307 identities, with each identity having more than 15 images. The model has 86% identification accuracy on the test set.

4.2. Evaluation Metrics and Comparison Methods

This paper follows the same evaluation metrics as DCNE [15] in quantitative experiments. For different XAI methods, we calculate the Intersection-over-Union (IoU) between the attribution map obtained by the method and the feature mask in the expert-annotated dataset, and find the attribution map with the largest IoU for each feature mask.Subsequently, we calculate the Mean IoU by averaging the IoU values across multiple samples of each bird species. This metric serves as an indicator of the alignment between the explanations provided by the XAI methods and human expert annotations, thereby assessing the completeness and comprehensibility of the explanations.

We compare the mean IoU scores of the features found by our method with CRP [1], Fullgrad [26], Gradcam [23], Gradcam++ [4], Layercam [13], Hirescam [5], XGradCAM [7] and DCNE [15]. Since some methods only generate one attribution map, in order to make the comparison more effective, we generate attribution maps by conditioning each convolutional layer in the network. Since the complexity of the explanation of CRP is relatively high, we construct CRP-10 using the attribution maps of the top 10 neurons (as the simplest concept reduction method). Since all concept reduction methods reduce the explanation of CRP, we use CRP-300 as the baseline for all concept reduction methods to show the best explanation quality that can be achieved with extremely high complexity.

Table 1. Mean IoU of XAI methods on expert-annotated dataset

Method	Crested Auklet					Parakeet Auklet					White crowned Sparrow			
	f-1 ↑	f-2 ↑	f-3 ↑	f-4 ↑	Avg ↑	f-5 ↑	f-6 ↑	f-7 ↑	f-8 ↑	Avg ↑	f-9 ↑	f-10 ↑	f-11 ↑	Avg ↑
GradCAM[23]	0.286	0.396	0.150	0.172	0.251	0.159	0.258	0.270	0.064	0.188	0.413	0.429	0.175	0.339
HiresCAM[5]	0.204	0.257	0.062	0.062	0.146	0.571	0.188	0.242	0.274	0.319	0.201	0.216	0.295	0.237
GradCAM++[4]	0.230	0.445	**0.264**	0.126	0.266	0.146	0.137	0.277	0.107	0.167	0.327	0.314	0.415	0.352
XGradCAM[7]	0.536	0.309	0.090	0.110	0.261	0.088	0.376	**0.536**	0.187	0.297	0.334	0.341	0.269	0.315
LayerCAM[13]	0.673	0.555	0.104	0.045	0.344	0.528	0.398	0.296	0.187	0.352	0.410	0.384	0.116	0.303
FullGrad[26]	0.386	0.206	0.046	0.014	0.163	0.482	0.269	0.313	0.073	0.284	0.220	0.261	0.142	0.208
CRP-10[1]	0.381	0.353	0.068	0.062	0.216	0.447	0.364	0.181	0.102	0.274	0.333	0.351	0.100	0.261
DCNE[15]	0.732	0.596	0.185	**0.195**	0.427	0.537	**0.739**	0.382	0.237	0.474	**0.642**	0.623	**0.486**	0.584
Ours	**0.750**	**0.639**	0.212	0.164	**0.441**	**0.577**	0.594	0.446	**0.314**	**0.483**	0.636	**0.671**	0.466	**0.591**

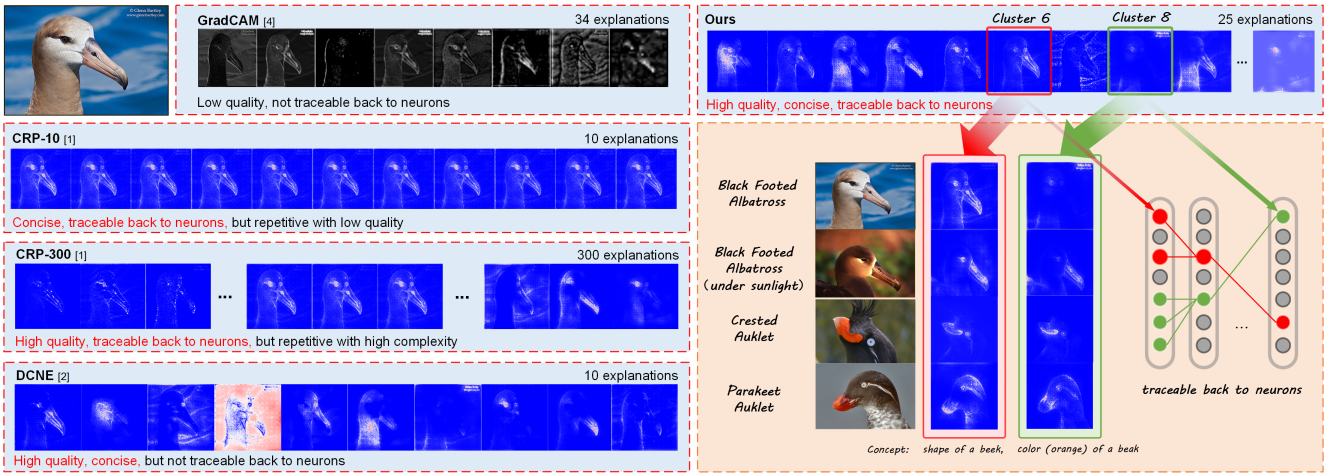

Figure 4. Comparison of explanations obtained by various methods. Our method obtains explanations that are of high quality, concise and can be traced back to neurons. Traceable explanations can provide a more nuanced interpretation of the model: *e.g.* a bird in sunlight may be incorrectly identified as different species with an 'orange beak', as the neurons encoding the concept of "orange beak" respond significantly in sunlight.

In addition, we compare the number of attribution maps produced by each method as a metric of explanation complexity with the method's average IoU over all samples. This comparison shows the trade-off between the quality of the explanation and the complexity of the explanation for each method.

4.3. Quantitative Experiment

Table.1 shows the performance of various explainable methods on a bird data set annotated by experts. Each sub-table corresponds to a specific bird species. Each column of the sub-table represents a feature that was annotated by experts and considered important to distinguish this species. For example, the annotated 'f-1' for the bird species 'Crested Auklet' is 'pale eye'. 'Avg' indicates the average of all feature IoU values. In addition, for ease of viewing, we highlight the value of the maximum mean IoU in each feature in bold, and the second mean IoU value is underlined.

As shown in Table.1, in most cases, our method achieves the best results in the average IoU of each feature of the several birds listed and achieves a higher explanation quality for each feature (obtaining the Mean IoU of the top two). Specifically, for the 'Crested Auklet' annotated feature 'f-1' ('pale eye'), our method achieved an average IoU value of 0.75, significantly higher than other methods. Additionally, our method performed exceptionally well on multiple features of the 'Parakeet Auklet', 'White-crowned Sparrow' and other bird species, with average IoU values consistently ranking in the top two. This consistency underscores the strong explanatory capability of our method across different bird species and features, effectively capturing key characteristics annotated by experts.

According to [1], CRP has a greater chance of finding an attribution map similar to the expert's annotation because it can access more feature maps. If CRP has 300 concepts, it can achieve a higher performance of average IoU of 0.632 for the case "White crowned Sparrow". However, a large number of feature maps also greatly increases the complexity of the explanation, making the concept incomprehensi-

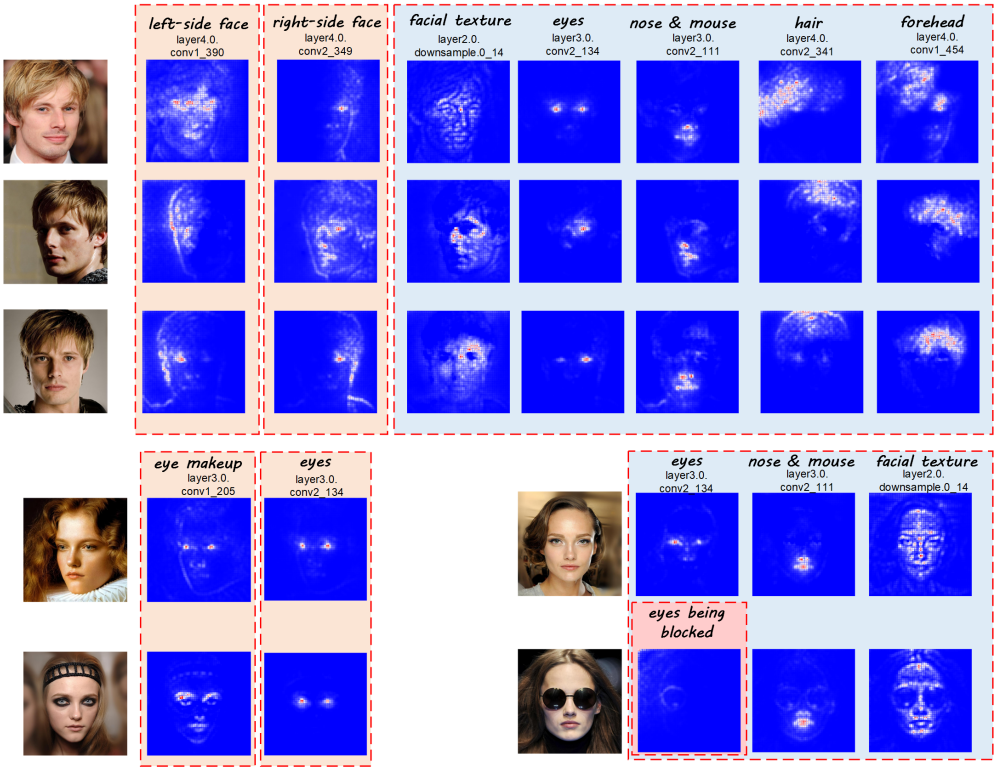

Figure 5. Explainable analysis of face recognition tasks using our method. Our method can provide a concise view of several concepts that are most important for identity recognition, such as 'facial texture', 'eyes', 'nose and mouth', 'hair', and 'forehead'. Besides, our method can trace back to some specific neurons attributed to the 'left-side face' and 'right-side face', which can be further studied for large pose face recognition. Concepts related to "eye makeup" can be found to explore the effect of makeup on face recognition. The occlusion of eyes such as "glasses" can be discovered by looking at the lack of concepts in the "eyes" clustering.

ble and unreadable. In contrast, our method achieves competitive performance with significantly more concise explanations.

Figure 3 presents a comparison of the mean IoU values and the explanation complexity of various methods. It can be observed that CRP-300 achieves the highest explanation quality, but at the cost of extremely high complexity. DCNE obtains considerable explanation quality with the lowest explanation complexity. In actual experiments, our method generates about 25 attribution maps, which obtains a complexity lower than that of methods other than DCNE, while achieving performance close to CRP-300. This validates that our method can reduce the complexity of the explanation while improving the quality of the explanation.

Overall, the effectiveness of the concept-neuron connection mechanism in Section 3.3 is validated in the quantitative evaluation. Our method achieves performance comparable to the CRP-300 method with only 25 concepts, primarily due to the concept-neuron connection mechanism effectively preserving key discriminative features, while the weighted distance metric effectively captures semantic relevance among neurons, reducing redundancy in the CRP-300 method. It is worth noting that when the number of explanations is similar (e.g., $ExplanationComplexity < 50$), our method shows a significant improvement in IoU compared to other methods. This is mainly due to the weight adjustment mechanism in Section 3.2 that optimizes the fusion ratio of shallow and deep features, as well as the automatic clustering guided by the silhouette coefficient that ensures an optimal balance between interpretation completeness and conciseness. In addition, the performance of our method still does not perfectly restore the interpretive integrity of CRP-300, which implies that "redundancy" is not completely ineffective. Although very small, "redundancy" does contribute to the interpretive integrity, which will be explored in future studies.

4.4. Qualitative Experiment

In the qualitative experiments, we first visualize the comparison of the explanations obtained by different explainable methods. As shown in Figure 4, we compare the traditional post-hoc explainable methods (with GradCAM [23] as a sample), CRP-10 [1], CRP-300 [1], DCNE [15], and our method. We can see that our method obtains high-quality, concise, and neuron-traceable explanations when explaining the same model for the same input image for the

classification task.

Each cluster in our approach represents a specific group of neurons. That is, a specific group of neurons always encodes a 'fixed' concept, even if this concept may not be important for the current task. For example, in Figure 4, 'Black footed Albatross' is a bird with a whitish body color and a pale beak, but in the context of 'sunset' it appears darker and reddish in body and beak. We can see from the clustering attribution map that a group of neurons representing 'a specific beak shape' responds to both lighting conditions, but another new group of neurons only responds significantly in the 'sunset' environment. We found that this particular new group of neurons is used to encode the 'orange beaks' of birds such as the 'Crested Auklet' and the 'Parakeet Auklet'. Knowing the specific neuron cluster that encodes the concrete concept can provide us with the possibility to interpret the network. In this case, we provide some evidence to explain why, under certain specific lighting conditions, 'Black Footed Albatross' is likely to be recognized as a bird with an orange beak similar to 'Crested Auklet' and 'Parakeet Auklet'.

Overall, we demonstrate the ability of our approach to understand the mechanisms within neural networks that process complex image information, effectively decompose complex network responses into explainable neuron clusters, thereby refining the understanding of model behavior.

4.5. Face Recognition Qualitative Experiment

In our further research, we extend our explainable method to face recognition to investigate neuronal responses to facial features and external factors. This section demonstrates how neural networks internally process face images and reveals valuable insights about the face recognition model through detailed analysis of neuron activation patterns.

Based on a ResNet-18 model [19] trained on the CelebAMask-HQ dataset [16], we conducted an explainable analysis of a facial recognition task. Our approach provides a concise view of several concepts that are most important for identity recognition, such as 'facial texture', 'eyes', 'nose and mouth', 'hair', and 'forehead' as shown in Figure 5.

Focus on New Concepts when Large Pose Face Recognition. It is worth noting that our method traces certain neurons attributed to the concepts of 'left-side face' and 'right-side face', providing assistance in the identification task. We analyzed the behavior of pose-tuning neurons. We found that when the input sample is a large-pose face and the poses match, the neurons are significantly activated. When the poses do not match, only the corresponding contours are activated. Such results can help us better understand how the model handles face images from different angles.

Robustness to Eye Makeup. As can be seen in Figure 5, when the network processes a face image with eye makeup, a specific set of neurons responds significantly more to the region near the "eyes". To eliminate the interference, we compared the neurons encoding the concept of 'eyes', and we can see that the group of neurons encoding the concept of 'eye makeup' does not respond strongly when there is little or no eye makeup. This finding suggests the possibility that the network encodes the concept of "eye makeup" to improve model accuracy.

Lack of attention to the concept of "eye occlusion". Occlusions of eyes such as "wearing glasses" can be detected by looking at the missing concepts in the "eyes" clusters. Figure 5 shows that in the case of "wearing glasses", the activity of some neurons that responded to the concept of "eyes" was significantly reduced. In addition, our approach does not find that the model encoded concepts related to "glasses". This is due to the small number of glasses-wearing samples in the original training data, and the model did not assign specific neurons to encode this concept.

5. Conclusion

In this paper, we propose a concept-based explainable method that identifies a set of explanations that are critical for classification through weighted hierarchical clustering. Our study contains the following three innovations. First, we significantly reduce the number of generated explanations, avoid redundant information, and make the explanations more concise and clear. Compared to traditional methods, our method significantly reduces the number of explanations while ensuring high quality. Second, we experimentally demonstrate that our method improves the completeness and comprehensibility of the explanations. Finally, in terms of providing new research perspectives, tracing explanations back to neurons provides researchers with new perspectives to further interpret and understand the decision-making process of the model. This approach shows significant potential for application in areas where transparency is required. We hope that this work will provide insights into the understanding of neural networks and provide inspiration for further model improvement.

Acknowledgement

This work was supported in part by Chinese National Natural Science Foundation Projects U23B2054, 62276254, 62306313, the Beijing Science and Technology Plan Project Z231100005923033, Beijing Natural Science Foundation L221013, the Science and Technology Development Fund of Macau Project 0140/2024/AGJ, and InnoHK program.

References

[1] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023.

[2] A. Akula, S. Wang, and S.-C. Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2594–2601, 2020.

[3] K. H. R. Chan, Y. Yu, C. You, H. Qi, J. Wright, and Y. Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114):1–103, 2022.

[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[5] R. L. Draelos and L. Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

[6] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023.

[7] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[8] S. Ghalebikesabi, L. Ter-Minassian, K. DiazOrdaz, and C. C. Holmes. On locality of local explanation models. *Advances in neural information processing systems*, 34:18395–18407, 2021.

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

[10] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[11] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[14] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[15] N. Kondapaneni, M. Marks, O. Mac Aodha, and P. Perona. Less is more: Discovering concise network explanations. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

[16] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[18] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[19] D. Na, S. Ji, and J. Kim. Unrestricted black-box adversarial attack using gan with limited queries. In *European Conference on Computer Vision*, pages 467–482. Springer, 2022.

[20] P. C. Neto, T. Goncalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso. Causality-inspired taxonomy for explainable artificial intelligence. *arXiv preprint arXiv:2208.09500*, 2022.

[21] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, and B. Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021.

[22] M. Scott, L. Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[24] K. R. Shahapure and C. Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.

[25] S. Sharma, N. Batra, et al. Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 568–573. IEEE, 2019.

[26] S. Srinivas and F. Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[27] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[29] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference,*

NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8, pages 563–574. Springer, 2019.

[30] J. Yang, X. Li, D. Pai, Y. Zhou, Y. Ma, Y. Yu, and C. Xie. Scaling white-box transformers for vision. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 36995–37019. Curran Associates, Inc., 2024.

[31] Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, B. Haeffele, and Y. Ma. White-box transformers via sparse rate reduction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 9422–9457. Curran Associates, Inc., 2023.

[32] H. Zhang, M. Xue, X. Liu, K. Chen, J. Song, and M. Song. Schema inference for interpretable image classification. *arXiv preprint arXiv:2303.06635*, 2023.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.