Probing Human Visual Strategies Using Explainability Methods for Artificial Neural Networks

Yousif Kashef Al-Ghetaa

A Thesis Submitted to the Faculty of Graduate Studies
In Fulfilment of the Requirements
For the Degree of Masters of Science

Graduate Program in Biology
York University
Toronto, Ontario

August 2024

Abstract

This thesis explores the explainability of artificial neural networks (ANNs) through the use of explainable artificial intelligence (XAI), specifically focusing on models of visual object recognition and their alignment with human visual processing. Given the increasing integration of ANNs in critical applications, understanding and improving the transparency of these models is paramount.

The research begins with a detailed comparative analysis of various ANN models, employing XAI techniques to generate explanations for their decision-making processes in visual recognition tasks. This study introduces the Sharpness Metric, a novel quantitative measure designed to assess how different explanations vary and which tools are best for distinguishing between them.

Further, the thesis examines how closely these machine-generated explanations align with human visual strategies. It employs innovative behavioral proxies to compare the effectiveness of ANN models against human cognitive processes, without the need for direct access to the internal workings of the human brain. This comparison is crucial for evaluating the potential of ANNs to mimic human-like reasoning in visual tasks.

The findings from these analyses are discussed in depth, highlighting the implications for the development of AI systems that are both interpretable and aligned with human cognitive processes. The thesis concludes by emphasizing the importance of these advances for real-world AI applications, such as autonomous driving and medical diagnostics, where making AI decisions understandable to humans is critical.

This work contributes to the fields of cognitive science and artificial intelligence by advancing our understanding of how ANNs can be made more transparent and how their operations can be more closely aligned with human visual processing. It opens up new pathways for research into creating more reliable, understandable, and aligned ANN models of primate vision.

Acknowledgments

I would like to extend my heartfelt thanks to Dr. Kar for welcoming me into his lab and providing invaluable mentorship throughout my time at York University. His dedicated guidance and support over the past year have been pivotal in advancing my academic achievements and have inspired me to engage deeply with my research.

Dr. Kar's profound passion, dedication, and expertise in visual neuroscience have been truly inspiring. I am eager to continue learning from him and further my understanding of this dynamic field.

Lastly, I am grateful to York University and the Department of Biology for giving me the opportunity to undertake this research and for the enriching experience it has provided.

List of Figures

Table of Contents

Chapter 1

Introduction

1.1 Models of Object Recognition

Human decision-making and object recognition are complex processes that have been studied extensively in psychology and neuroscience. Understanding how humans make decisions and recognize objects can provide valuable insights into the cognitive processes underlying vision. Researchers in psychology have proposed various models to explain these processes, such as the prototype model [Rosch et al., 1976] and the exemplar model [Medin and Schaffer, 1978]. The prototype model suggests that we compare new objects to an average or idealized representation of a category, while the exemplar model suggests that we compare new objects to specific examples we have encountered in the past. Both models have been supported by experimental evidence, and researchers continue to refine and explore these theories [Smith, 2014]. In addition to these models, artificial neural networks such as Convolutional Neural Networks (CNNs) [Krizhevsky et al., 2012a] [He et al., 2016], and vision transformers [Dosovitskiy et al., 2020] have emerged as powerful tools for image classification and object recognition in machine learning. Interestingly, they have also been shown to be effective models of primate visual processing [Yamins and DiCarlo, 2016]. This is not surprising, given that CNN architectures have been motivated by our understanding of the visual cortex in primates – organized in a hierarchical manner, with each level extracting increasingly complex features from the visual input [Hubel and Wiesel, 1962].

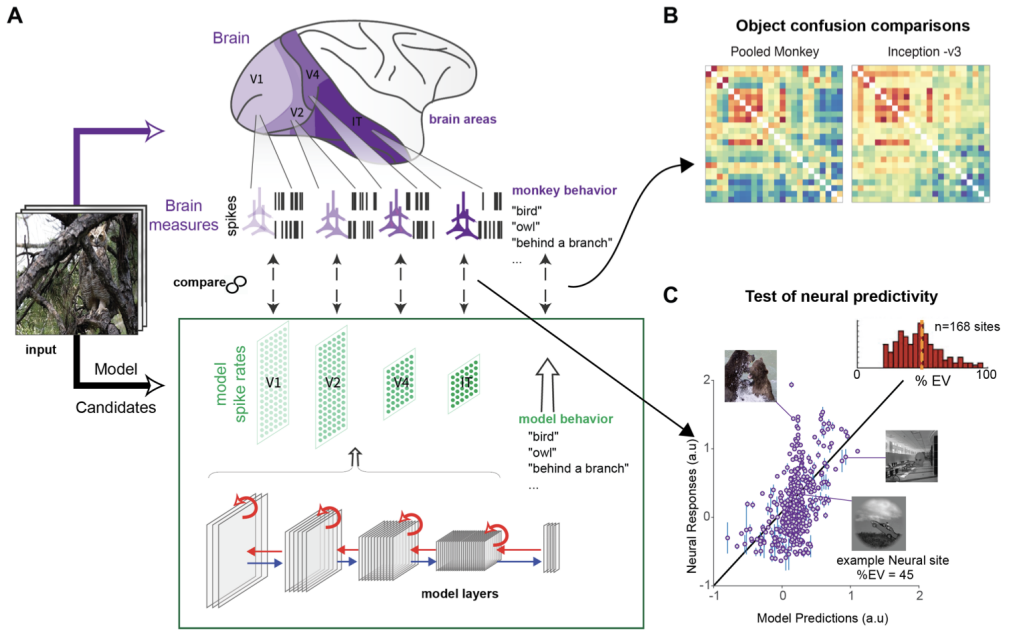

Figure 1.1: This figure depicts the methodology and evaluation of biologically plausible models of human vision based on neural recordings. Panel A outlines the procedure, starting with an input image that is processed by a computational model to generate candidate visual representations. These are compared against actual brain measures from various visual areas (such as V1, V2, V4, and IT) represented by spike patterns, aiming to simulate human object recognition behaviors, like identifying a bird or an object behind a branch. Panel B presents a comparison of object confusion between the aggregated monkey data and the computational model, reflecting the likelihood of misidentification of objects. Panel C offers a scatter plot correlating the model's predictions with neural responses, supported by a histogram detailing the percentage of explained variance in the neural sites. The combined data suggest how closely the model's predictions align with biological responses, thereby gauging the model's biological plausibility in replicating aspects of human visual perception. Figure adapted from [Kar and DiCarlo, 2024]

CNNs are also hierarchical in nature, with multiple layers that extract

increasingly complex features from the input image. This shared organization has led to the suggestion that CNNs may serve as a useful model for studying the neural basis of visual perception in primates. In fact, several studies [Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014] have shown that CNNs optimized to produce high accuracies in large-scale object recognition tasks can predict neural responses in the primate visual cortex. Thus, investigating the correlation between primates' visual recognition capabilities and machine learning models, such as CNNs, presents significant opportunities for enhancing our knowledge of primate vision. This includes examining the visual strategies that primates use in object recognition tasks.

Figure 1.1 demonstrates some methodologies human vision is modelled using a variety of tools. Panel A of Figure 1.1 outlines a systematic approach where an input image is processed through a computational model to generate candidate visual representations. These representations are then compared against neural recordings from various visual areas in the brain, such as V1, V2, V4, and IT, which are measured by spike patterns. This comparison illustrates how an ANN can simulate human object recognition behaviors, such as identifying a bird or discerning an object behind a branch. In Panel B, the figure shows a comparison of object confusion between aggregated monkey data and the computational model, highlighting the model's ability to reflect the likelihood of misidentifying objects. Panel C further illustrates the correlation between the model's predictions and neural responses, supported by a scatter plot that details the percentage of explained variance in neural sites. The combined data from these panels suggest that the model's predictions align closely with biological responses, thereby evaluating the model's biological plausibility in replicating aspects of human visual perception. This figure underscores the importance of integrating computational models with empirical neural data to advance our understanding of object recognition processes in the brain. While alignment between models of the ventral stream and actual neural activations and behaviors has been achieved, the congruence in terms of underlying visual strategies is yet to be thoroughly investigated.

Given the significance of aligning computational models with biological vision, ANNs present a compelling choice. ANNs not only achieve high accuracy in visual tasks but also allow for a detailed investigation of their underlying strategies using established methods (As opposed to biological systems). Techniques such as feature visualization, attribution methods, and network dissection enable researchers to dissect and interpret the features learned

by ANNs, offering insights into how these models mimic the human visual processing pathways. By leveraging these methods, we can thoroughly analyze the strategies employed by ANNs, providing a deeper understanding of their alignment with human vision and enhancing their biological plausibility. While there is as variety of ANN architectures in the field of machine learning, such as transformers, with superhuman performance, We chose to rely on Convolutional Neural Networks (CNNs) in our study for their high degree of neural and behavioral alignment with biological vision [Schrimpf and Prescott-Roy].

1.2 Explainable Artificial Intelligence for Models of Primate Vision

The explainability of Artificial Neural Networks (ANNs) has become a crucial topic in the field of machine learning and artificial intelligence. As ANNs, particularly deep learning models, have achieved state-of-the-art performance in various tasks, understanding their decision-making processes and internal representations is essential for several reasons. First, explainability can help build trust in the predictions made by these models, especially in high-stakes applications such as medical diagnosis, finance, and autonomous vehicles [Ribeiro et al., 2016]. When users can comprehend and justify the decisions made by ANNs, they are more likely to trust and adopt these models in real-world applications. Second, explainability can lead to the discovery of novel insights and knowledge hidden within the learned representations of ANNs, which can be useful for refining existing theories or generating new hypotheses in various domains [Olah et al., 2018]. For instance, in the context of primate vision, understanding the internal workings of CNNs can provide valuable insights into the neural mechanisms that underlie visual perception. Third, improving explainability can help identify and mitigate potential biases in the training data or model architecture, ensuring fairness and preventing discrimination [Bar, 2019]. As ANNs are increasingly used in decision-making processes that affect people's lives, understanding their inner workings becomes necessary to ensure that these models do not make wrong or unethical decisions. furthermore, understanding AI decisions is starting to become a legal requiremnt in some juristicitions such as the European Union [Commission, 2021] Fourth, explainability could have intriguing engi-

neering applications, enhancing efficiency and performance in technological solutions.

For example, in the development of Apple's Vision Pro headset [app, 2024], the biological concept of human foveated vision was applied. This approach enables the headset to render less of the scene while maintaining a high refresh rate, optimizing the user's visual experience without overburdening the system's processing capabilities. Building on this concept, a further optimization could involve rendering only the parts of objects that humans most rely on for recognition. This selective rendering technique could dramatically reduce computational requirements and energy consumption, making devices more efficient and responsive. Such advances underscore the practical benefits of interpreting and applying insights from ANN's decision-making processes and internal representations, bridging the gap between artificial intelligence and human-centric engineering solutions.

The explainability of ANNs is critical for fostering trust, gaining insights, and ensuring fairness in their applications. To elucidate the intricate decision-making processes of these models, researchers have been developing various techniques and methodologies aimed at enhancing their explainability. However, a key challenge lies in determining the reliability and effectiveness of these methods. Many of these approaches are grounded in mathematical and psychological principles that may not directly contribute to the goal of clarification [A. and R., 2023]. Consequently, there is an ambiguity regarding which methods are genuinely interpretable to human observers. It is imperative, therefore, to establish criteria for evaluating these methods. While some techniques may indeed offer valuable insights into the decision-making processes of both humans and ANNs, others may merely constitute complex mathematical exercises with little relevance to the ultimate objective of facilitating human understanding of ANN decision-making processes. The primary issue is that the features and patterns identified and highlighted by some explanation methods may not always align with human intuition or understanding. Moreover, the 'disagreement problem', which underscores the significant issue that the diversity of XAI methods can yield divergent explanations for identical model outputs from the same image input. This phenomenon not only highlights the complexity inherent in XAI methodologies but also raises critical questions about the reliability and consistency of explanations generated by different XAI approaches. [Krishna et al., 2024]. While it is anticipated that different techniques will yield distinct perspectives, a critical question emerges: which of these explanations is most aligned

with the human, particularly the primate, visual system?

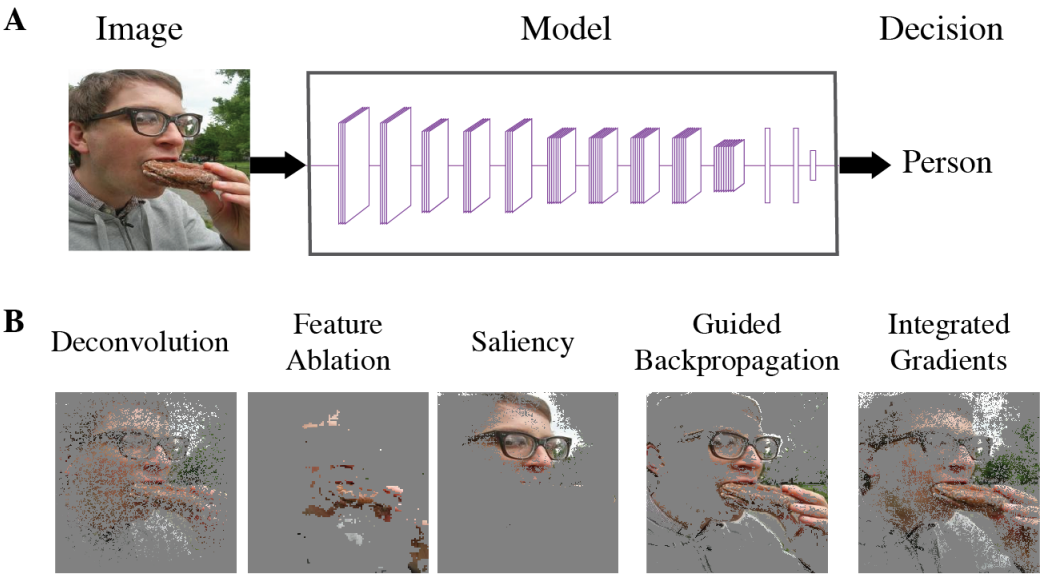

Figure 1.2: **The problem: XAI tools produce diverse outputs for the same stimuli, model and task. A.** An ANN model receives an image of a person as an input, and is able to assign the correct label. **B.** Various explanation methods (such as Deconvolution [Zeiler and Fergus, 2014], Feature Ablation, Saliency [Simonyan et al., 2013], Guided Backpropagation [Springenberg et al., 2014], Integrated Gradients [Sundararajan et al., 2017b]) yield diverse outputs for the example above.

Figure 1.2 illustrates the challenge of achieving explainability in ANNs through various XAI techniques. In Panel A, we see an ANN model tasked with recognizing and labeling a person from an image. The model processes the image and outputs a decision, identifying the subject in the image as a person. This straightforward process highlights the model's functionality but not the complexity of its decision-making process.

Panel B dives deeper into the complexity of model explainability by presenting outputs from four different XAI methods: Deconvolution [Zeiler and Fergus, 2014], Feature Ablation [Simonyan et al., 2013], Saliency, and Guided Backpropagation [Springenberg et al., 2014], along with Integrated Gradients. Each method aims to highlight the features and patterns within the

image that were most influential in the model's decision-making process. However, the diversity in the outputs illustrates a significant challenge in the field of XAI: each method provides a unique perspective on what the model deems important, leading to varying interpretations and potentially conflicting explanations of the same decision. Panel B dives deeper into the complexity of model explainability by presenting outputs from four different XAI methods: Deconvolution [Zeiler and Fergus, 2014], Feature Ablation [Hameed et al., 2022] , Saliency , and Guided Backpropagation, along with Integrated Gradients [Sundararajan et al., 2017b]. Each method aims to highlight the features and patterns within the image that were most influential in the model's decision-making process. However, the diversity in the outputs illustrates a significant challenge in the field of XAI: each method provides a unique perspective on what the model deems important, leading to varying interpretations and potentially conflicting explanations of the same decision.

Here, we propose developing a unified framework to compare visual recognition decisions made by primate and machine vision systems, aiming to functionally and structurally align computer vision models with models of primate vision. The research will focus on evaluating various methods of generating explanations to understand the differences between explanations generated by primate and machine learning models.

1.3 Limitations of Explainability Methods in Human Decision-Making

In the pursuit of understanding the visual strategies humans use in visual recognition, methodologies such as the Bubbles technique Gosselin and Schyns [2001] and Classification Images [Ahumada, 1996] have emerged as pivotal tools within the field of psychophysics. These methods have been standard for generating feature importance maps of the parts of the image relied on by humans.

The Bubbles technique, through its approach of selectively unveiling portions of an image, allows for a granular analysis of feature importance. This method facilitates a deeper understanding of the hierarchical processing in human vision, pinpointing the essential elements that contribute to the recognition of complex images. Conversely, Classification Images adopt a statisti-

cal lens, amalgamating data across numerous trials to distill the pixel-level contributions to perceptual tasks. This aggregation unveils the subtle, yet critical, visual cues that guide human perception, providing a composite image that highlights the pixels most influential to decision-making processes.

More recent advances in psychophysics, such as the Generalized Linear Model (GLM) [Murray, 2017], have aimed to illustrate the consistency and interconnectedness of visual strategies. The GLM successfully integrates classification and bubbles images within a single framework, underscoring their similarities and enabling concurrent measurements. However, the stimuli used to estimate bubbles and classification images using GLM techniques were very simple and the advantages of using such statistical methods may not generalize well to the complex natural scenes typically used for object recognition tasks.

Despite their contributions to the field, these methodologies are not devoid of limitations. The Bubbles technique, for instance, may inadvertently overlook crucial interactions between visual elements [Murray and Gold, 2004]. Its discrete revelation of image features can lead to a fragmented understanding of object recognition, as it fails to account for the Gestalt principles that govern human perception, where the whole is often perceived as more than the sum of its parts [Wertheimer, 1912]. This oversight can impede the accurate identification of interdependent features within an image, thereby skewing the interpretation of feature importance.

On the other hand, the Classification Images method grapples with the introduction of noise artifacts. The very nature of aggregating data from multiple trials, each with its inherent variability, can result in the amplification of noise. This accumulation of extraneous data can obscure the true signal, manifesting as noise artifacts within the generated importance maps [Murray, 2011]. Such artifacts not only challenge the clarity of the results but also pose the risk of misinterpretation, as they may falsely suggest the significance of irrelevant image regions. The delicate balance between signal clarity and noise interference becomes a pivotal concern in the application of Classification Images, necessitating rigorous methodological refinements to mitigate these drawbacks.

While the Bubbles technique and Classification Images stand as cornerstone methodologies in psychophysics for elucidating the underpinnings of human visual perception, their efficacy is tempered by inherent limitations. The Bubbles technique's potential oversight of interdependent visual elements and the Classification Images method's susceptibility to noise artifacts

underscore the necessity for further methodological advancement. Addressing these challenges is paramount for the refinement of these techniques, ensuring their ongoing contribution to our understanding of the ccomplexity of human perception.

The elucidation of the *visual strategies* employed by the primate ventral stream during visual decision-making processes, as well as the understanding of these strategies in models of the primate ventral stream, continues to present a significant challenge. Despite successful alignment of behavioral data with neural activation patterns, the critical task remains to identify the salient features upon which these systems depend. Furthermore, it is imperative to determine whether models of the ventral stream similarly utilize these salient features. Our investigations have been constrained by the limitations inherent in the psychophysical techniques previously discussed. Consequently, there is a pressing need to develop innovative methodologies for the estimation of primate saliency maps, which could offer new insights into the underlying mechanisms of visual processing in primates. In the field of machine learning and artificial intelligence, model explanation refers to the process of understanding and interpreting how a specific machine learning model makes predictions or decisions [Ribeiro et al., 2016]. Model explanation is crucial because, as models become more complex, comprehending the rationale behind their decisions becomes challenging, leading to a lack of transparency. This obscurity can impede trust in, or refinement of, the model. In contrast to the straightforward explainability of simpler predictive models such as linear regression—which owes its clarity to the limited number of parameters—deep learning models, characterized by their extensive parameter sets and diverse architectural configurations, present a significant challenge in this regard. Consequently, there is a pressing need to devise and implement explainability methodologies that can *demystify the abstractions and data processing procedures of deep learning models by contextualizing their operations in relation to the input data.*

By providing explanations for a model's predictions or decisions, we can glean insights into its functioning, identify potential biases or errors, and enhance the model's overall performance [Olah et al., 2018]. XAI methods can be broadly categorized into a few major families based on their underlying principles and methodologies. The following goes over the families and the specific methods we chose to employ in out studies. Figure 1.3 shows examples from all the explanation methods used in this study.

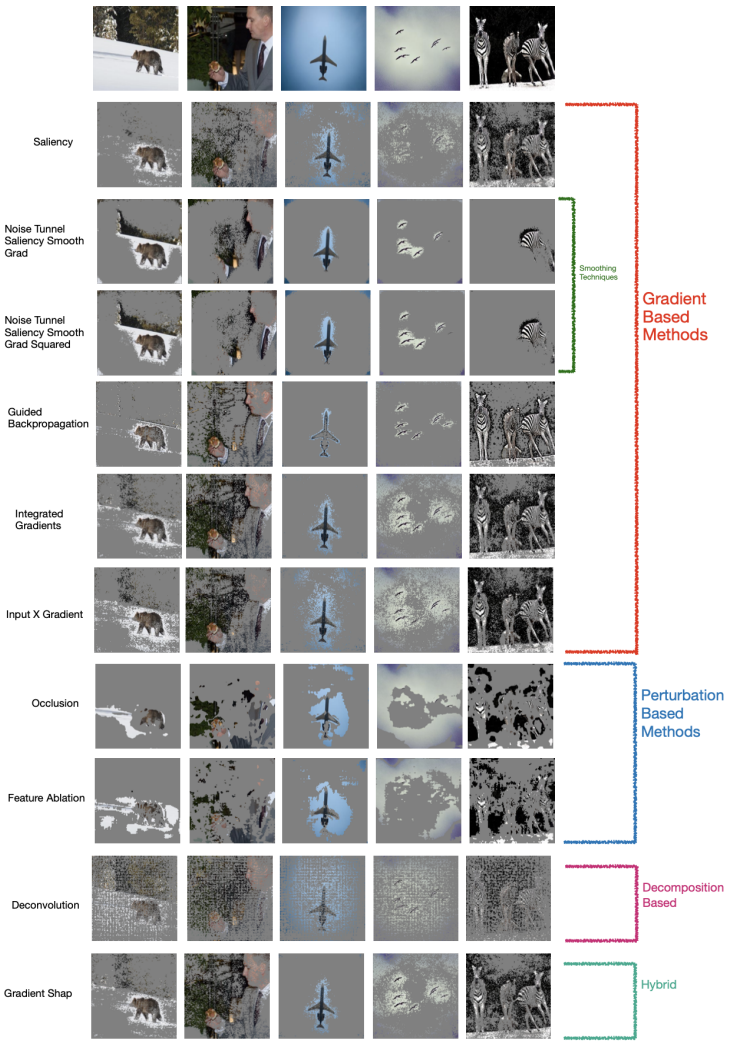

Figure 1.3: **Visualizations of Various XAI Methods Employed in the Study Using a Single Model**This figure displays various explanation methods categorized into Gradient-Based, Perturbation-Based, Decomposition-Based, and Hybrid Methods, applied using a single model. Each row demonstrates techniques like Saliency, Occlusion, and Integrated Gradients, showcasing how the model interprets visual information. These visualizations help understand the model's decision-making processes and improve its accuracy and fairness.

Gradient-based Methods

Techniques in this family calculate how changes in input features affect a model's output, using the concept of gradients or derivatives [Nielsen et al., 2022]. Notable methods in this category include:

Saliency:

Concept and Core Mechanism: The saliency method is a fundamental technique for visualizing which parts of an input image are most influential in a model's decision-making process. This is achieved by calculating the gradient of the model's output with respect to each pixel in the input image. Essentially, this involves computing the first derivative of the output with respect to the input image, resulting in a saliency map that highlights the pixels which, if altered, would most affect the model's prediction (also known as activation maximization) [Simonyan et al., 2013].

Technical Implementation: To generate a saliency map, we start by selecting a specific class c for which we want to understand the model's decision. Let $S_c(I)$ be the score of class c for a given image I. The saliency method computes the gradient of S_c with respect to the input image I. Formally, this is represented as $\frac{\partial S_c}{\partial I}$. This gradient highlights the regions in the image that the model finds most pertinent for its classification. These values are then visualized, usually by taking the absolute values of the gradients and creating a heat map over the image.

Visual Interpretation: Saliency maps offer a straightforward and intuitive way to understand what the model focuses on in an image. By examining these maps, one can see which areas of an image are deemed most important by the model for making a particular classification decision. This visual explanation is crucial for validating and interpreting the behavior of deep learning models, especially in critical applications like medical imaging or autonomous driving.

Biological Parallel: Although the computation of pixel-wise gradients is a mathematical operation that does not occur in biological vision, the concept of focusing on important regions is analogous to the biological process of foveation in human vision. In human vision, the eyes move to focus on areas of interest to gather detailed information, much like how a saliency map highlights important regions in an image for a model's decision-making process.

Applications and Benefits: Saliency maps can be used not only for understanding model predictions but also for tasks such as weakly supervised

object localization. By identifying the regions of an image that contribute most to the class score, these maps can help in segmenting objects within the image without needing detailed annotations.

Example and Practical Insights: For instance, in the context of a deep convolutional network trained on the ImageNet dataset, generating a saliency map for a particular class involves the following steps:

1. **Forward pass:** Compute the class score for the input image.
2. **Backward pass:** Calculate the gradient of the class score with respect to the input image.
3. **Visualization:** Create a heat map of the gradient magnitudes, highlighting the most influential pixels.

This method has been shown to be effective in various studies and is a valuable tool for researchers and practitioners aiming to interpret and trust the decisions made by deep learning models.

Guided Backpropagation:

Concept and Core Mechanism: Guided Backpropagation is an enhancement of the traditional backpropagation method used during model training [Springenberg et al., 2015]. It modifies the backpropagation process to only allow positive gradients to pass through, meaning it focuses on what actively enhances the prediction rather than what reduces it. This adjustment creates cleaner, more precise visualizations of the features that matter to the model but does omit features that suppress excitation.

Technical Implementation: In standard backpropagation, both positive and negative gradients are propagated through the network during the training process. Guided Backpropagation changes this by zeroing out negative gradients during the backward pass. Formally, if g is the gradient at a particular neuron, Guided Backpropagation modifies it such that:

$$g' = \begin{cases} g & \text{if } g > 0 \\ 0 & \text{otherwise} \end{cases}$$

This selective process ensures that only the elements contributing positively to the activation are considered, thus providing a more refined view of the input features that are driving the model's predictions.

Visual Interpretation: The result of Guided Backpropagation is a more detailed and clearer visualization of the input features that are important to the model. This is because it eliminates noise caused by negative contributions, thereby highlighting distinct features in the image that most strongly activate the model.

Biological Parallel: This selective enhancement can be loosely analogous to the brain's mechanism of reinforcing successful outcomes, where positive reinforcement strengthens certain neural pathways. However, it is important to note that the actual neural mechanisms are not directly comparable to the operations of Guided Backpropagation.

Applications and Benefits: Guided Backpropagation is particularly valuable for providing clearer insights into the model's behavior. By highlighting distinct features that are important for the model's decision, it allows for a better understanding and validation of the model, which is crucial in applications requiring high explainability, such as medical imaging and security.

Comparison with Saliency Maps: While both Guided Backpropagation and Saliency Maps aim to visualize important features in the input data, they differ in their approach. Saliency Maps compute the gradient of the class score with respect to the input image, resulting in a visualization that highlights all influential pixels, both positive and negative. In contrast, Guided Backpropagation only propagates positive gradients, which leads to cleaner and more precise visualizations by focusing on features that enhance the model's prediction.

Integrated Gradients:

Concept and Core Mechanism: Integrated Gradients offers a deeper look at feature importance by integrating the gradient along the path from a baseline to the actual input [Sundararajan et al., 2017b] This method not only considers the immediate gradient but accumulates this gradient over a series of steps from a non-informative input to the actual input. This accumulated gradient provides a thorough attribution of how each feature in the input contributes to the final prediction, making it robust against certain biases that simpler gradient methods might introduce.

Technical Implementation: To compute Integrated Gradients, we define a baseline input $\mathbf{I}_0$, which is typically a zero vector or some reference

input. The integrated gradient for feature i is calculated as:

$$\text{IntegratedGradient}_i(\mathbf{I}) = (\mathbf{I}_i - \mathbf{I}_{0_i}) \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{I}_0 + \alpha(\mathbf{I} - \mathbf{I}_0))}{\partial \mathbf{I}_i} d\alpha$$

where F is the model's output function, and α is a scalar that scales the input from the baseline to the actual input. This integral accumulates the gradients at points along the straight-line path from the baseline to the input, providing a comprehensive measure of each feature's contribution.

Visual Interpretation: Integrated Gradients result in an attribution map that shows how much each input feature contributes to the final prediction. Because it integrates over a range of inputs, it can highlight feature importance in a way that accounts for complex interactions between features, making it particularly useful for models with such characteristics.

Biological Parallel: Although humans do not compute gradients, the notion of considering a range of scenarios from a neutral baseline to the current state can be likened to human decision-making processes that evaluate different possibilities and their outcomes.

Applications and Benefits: Integrated Gradients is particularly useful for models where input features have complex interactions, as it can provide a detailed breakdown of how each feature contributes to the outcome over different input scenarios. This method is robust against certain biases that simpler gradient methods might introduce and provides a more reliable interpretation of feature importance. Integrated Gradients helps mitigate biases like feature dependence, where the importance of one feature is disproportionately assessed based on the values of other correlated features, and gradient saturation, where traditional methods may underestimate the contribution of features in regions where the model's predictions have plateaued.

Comparison with Saliency Maps and Guided Backpropagation: While Saliency Maps and Guided Backpropagation focus on immediate gradients, Integrated Gradients accumulates gradients over a range of inputs from a baseline to the actual input. Saliency Maps highlight all influential pixels, both positive and negative, whereas Guided Backpropagation zeroes out negative gradients to provide cleaner visualizations of enhancing features. Integrated Gradients, on the other hand, offers a more comprehensive view by integrating gradients along the input path, making it robust and capable of handling complex feature interactions.

Input X Gradient:

Concept and Core Mechanism: The Input X Gradient method multiplies the input by its gradient, emphasizing the interaction between the input value and its gradient [Shrikumar et al., 2017]. By doing so, it highlights areas of the input that would most affect the output if changed. This approach is straightforward yet powerful, as it directly ties the change in output to specific input features, providing clear insight into which parts of the input are most critical for the model's decisions.

Technical Implementation: To compute Input X Gradient, we take the element-wise product of the input $\mathbf{I}$ and the gradient of the model's output with respect to the input:

$$\text{InputXGradient}_i(\mathbf{I}) = \mathbf{I}_i \cdot \frac{\partial F(\mathbf{I})}{\partial \mathbf{I}_i}$$

where F is the model's output function, and $\mathbf{I}_i$ is the i-th feature of the input. This product directly correlates the input feature values with their corresponding gradients, highlighting the most influential features.

Visual Interpretation: The result of Input X Gradient is an attribution map that shows how much each input feature contributes to the model's output. This method provides a clear and intuitive understanding of feature importance by linking input values directly with their impact on the output.

Biological Parallel: This method's straightforward approach to feature importance does not parallel any known biological processes directly but simplifies understanding in a way that is accessible and logical, akin to basic cause-and-effect reasoning in human cognition.

Applications and Benefits: Input X Gradient is useful for gaining quick and clear insights into which parts of the input are most critical for the model's decisions. Its simplicity makes it a practical choice for many applications, providing a direct measure of feature importance.

Comparison with Saliency Maps, Guided Backpropagation, and Integrated Gradients: - textitSaliency Maps: Saliency Maps compute the gradient of the output with respect to the input image, highlighting influential pixels but not necessarily correlating their values with their impact. Input X Gradient, by contrast, directly ties input values with their gradients, providing a more immediate understanding of feature importance. - textit-Guided Backpropagation: Guided Backpropagation enhances the traditional backpropagation by allowing only positive gradients, resulting in cleaner visualizations. Input X Gradient, while also straightforward, does not filter gradients but instead multiplies them by the input values, offering a different

perspective on feature importance. - textitIntegrated Gradients: Integrated Gradients accumulate gradients over a range of inputs from a baseline to the actual input, providing a comprehensive view of feature importance. Input X Gradient is more immediate and less complex, focusing directly on the interaction between input values and their gradients without integrating over multiple inputs.

Perturbation-based Methods

Perturbation-based methods assess the importance of different features by altering them and observing how these changes affect the output. These methods are essential for models where direct gradient computation is not feasible.

Occlusion:

Concept and Core Mechanism: By systematically blocking out different parts of the input and observing how the model's predictions change, occlusion helps identify which areas are most crucial for the model's performance [Zeiler and Fergus, 2014]. This method involves masking portions of the input data (e.g., patches of an image) and then measuring the change in the model's output. The regions whose occlusion leads to significant changes in the prediction are deemed important.

Technical Implementation: To implement occlusion, the input image $\mathbf{I}$ is systematically occluded using a mask $\mathbf{M}$. The mask is moved across the image in a sliding window fashion, and at each position, the occluded input $\mathbf{I}_{\text{occluded}} = \mathbf{I} \odot (1 - \mathbf{M})$ is fed into the model. The change in the model's output $\Delta F = F(\mathbf{I}) - F(\mathbf{I}_{\text{occluded}})$ is recorded. This process is repeated for multiple positions of the mask, creating a heatmap that highlights the importance of different regions of the input.

Visual Interpretation: Occlusion provides a straightforward and intuitive visualization of feature importance. By observing how the model's prediction changes when different parts of the input are occluded, one can directly see which regions are vital for the model's decision-making process. The resulting heatmap shows the importance of different regions, with larger changes in prediction indicating more critical areas.

Biological Parallel: Biologically, this method can be related to psychophysical techniques such as the bubbles method, which involves revealing parts of an image through small windows to study perception. However, the

occlusion operation itself has no direct neural parallels.

Applications and Benefits: Occlusion is particularly effective for spatial data like images. It provides clear visual evidence of feature importance by simply removing parts of the input and noting the effect. This method is easy to understand and implement, making it a useful tool for interpreting and validating model predictions, especially in applications where spatial context is critical, such as medical imaging and autonomous driving.

Feature Ablation:

Concept and Core Mechanism: Similar to occlusion, feature ablation involves deliberately altering or removing features to assess their impact on the model's output [Hameed et al., 2022]. This method systematically removes or modifies individual features or sets of features in the input data to observe how the model's predictions change. The change in performance indicates the importance of the ablated features.Feature ablation in images involves altering or removing aspects such as color channels, textures, shapes, background elements, and spatial relationships to assess their impact on model outputs.

Technical Implementation: To perform feature ablation, the input $\mathbf{I}$ is modified by setting certain features $\mathbf{I}_{\text{ablated}}$ to zero or another baseline value. The model is then evaluated on this modified input, and the change in the model's output $\Delta F = F(\mathbf{I}) - F(\mathbf{I}_{\text{ablated}})$ is measured. By systematically ablating different features or combinations of features, one can create an importance map that highlights which features are most critical for the model's decisions.

Visual Interpretation: Feature ablation provides a clear view of feature importance by directly showing how the removal or alteration of specific features affects the model's predictions. This method helps in understanding the role of individual features and their contributions to the overall model output.

Applications and Benefits: Feature ablation is invaluable for understanding complex models where multiple features interact to influence the prediction. It helps in identifying redundant or irrelevant features, which can be critical for model simplification and interpretation. By highlighting the most impactful features, feature ablation aids in refining and improving model performance and explainability.

Comparison with Occlusion: - textitOcclusion: Occlusion is primarily used for spatial data, like images, and involves blocking out patches of

the input to see how the model's predictions change. It creates a heatmap showing the importance of different spatial regions. Occlusion is straightforward and intuitive, providing clear visual evidence of which regions are vital for the model's decision-making process. - textitFeature Ablation: Feature ablation, on the other hand, can be applied to any type of data and focuses on systematically removing or altering individual features or sets of features to assess their impact on the model's output. It provides detailed insights into the relevance of specific features, making it especially useful for models with complex feature interactions. Feature ablation helps identify redundant or irrelevant features, aiding in model simplification and enhancing explainability.

Decomposition Methods

Decomposition methods break down the model's decision-making into more understandable components by tracing decisions back to the input features.

Deconvolution:

Concept and Core Mechanism: Deconvolution is a method that reverses the model's computations to map its decisions back to the input space, identifying which features activate certain filters within convolutional neural networks (CNNs) [**?**]. This technique helps to trace the model's decision-making process by visualizing the activations at each layer of the network, effectively showing how different features contribute to the final output.

Technical Implementation: In a CNN, deconvolution involves reversing the forward pass to map the activations back to the input image. This is done by successively applying the transposed operations of the original convolutional layers. For example, the process includes the following steps:

1. **Unpooling**: Reverses the pooling operation to restore the spatial dimensions of the feature maps.

2. **Rectified Linear Unit (ReLU) Inversion**: Applies the ReLU function in reverse to retain only the positive activations.

3. **Inverse Convolution**: Uses the transposed convolution (also known as deconvolution) to map the feature activations back to the input space.

By applying these steps, deconvolution reconstructs the input image from the activations, highlighting the areas that strongly activate specific filters.

Visual Interpretation: Deconvolution provides a detailed visualization of the features that activate different filters within the CNN. It helps to understand the hierarchical nature of the features learned by the model, from low-level edges and textures to high-level object parts. This visualization is crucial for interpreting how each layer of the network processes and transforms the input data to make predictions.

Applications and Benefits: Deconvolution is particularly significant for visualizing and understanding models that use layers of filters, such as CNNs. It helps clarify how each layer transforms the input to arrive at a decision, providing insights into the inner workings of the network. This method is valuable for model validation, debugging, and enhancing the explainability of deep learning models, especially in fields like computer vision where understanding the learned features is essential. Deconvolution is used in facial recognition to identify key facial features, in autonomous vehicles to interpret road scenes, in medical imaging to pinpoint anomalies like tumors, and in surveillance to distinguish activities, thereby enhancing model interpretability and reliability in various applications.

Hybrid Methods

Hybrid methods blend principles from various families to exploit their strengths while mitigating their weaknesses.

Gradient SHAP:

Concept and Core Mechanism: Gradient SHAP combines SHAP (SHapley Additive exPlanations) values, which quantify the impact of each feature by comparing different possible combinations, with gradient information to provide a detailed attribution of feature importance [Lundberg and Lee, 2017]. This method leverages the strengths of both SHAP values and gradient-based attributions to capture the influence of each feature across different conditions, offering a nuanced understanding of feature contributions.

Technical Implementation: Gradient SHAP involves the following steps:

1. **Baseline Distribution**: It starts by defining a distribution of baselines, which are typically samples of the input with some features set

to reference values (e.g., zero).

2. **Gradient Calculation**: For each baseline, the gradient of the model's output with respect to the input is calculated.

3. **SHAP Value Integration**: The SHAP values are integrated with these gradients to compute the attribution of each feature. This integration is done by averaging the gradients across multiple baseline samples, thus providing a comprehensive view of feature importance.

The resulting attributions reflect both the gradient information and the distribution of feature impacts across different possible input combinations.

Visual Interpretation: Gradient SHAP provides a detailed visualization of feature importance by combining gradient-based attributions with SHAP values. This method highlights how each feature contributes to the model's output under various input conditions, resulting in a nuanced and comprehensive attribution map.

Applications and Benefits: Gradient SHAP excels in both transparency and accuracy, making it an essential tool for complex models where individual feature effects are difficult to isolate. It is particularly useful in scenarios where understanding the interaction and combined effects of multiple features is crucial, such as in finance, healthcare, and other domains with intricate data relationships.

Biological Parallel: The integration of multiple analytical methods in Gradient SHAP reflects complex human problem-solving, where multiple perspectives and methods are combined to understand a problem better. However, the computational specifics of Gradient SHAP do not have direct biological counterparts and do not parallel any known mechanisms in the human ventral stream, which is involved in object recognition and visual processing.

Smoothing Techniques

These techniques focus on enhancing the explainability of saliency maps by integrating results over multiple instances or adding noise to reduce variability.

- **Noise Tunnel with Saliency (SmoothGrad)**: By adding noise to the input and averaging the results of multiple saliency maps, SmoothGrad aims to provide a clearer and more consistent visualization of

feature importance [Adebayo et al., 2020]. This method is crucial for reducing the noise inherent in single saliency maps, providing a more stable and reliable view of what features the model deems important.

- **Noise Tunnel with Saliency (SmoothGrad-Squared)**: Building on SmoothGrad, this method squares the gradients before averaging, which amplifies consistent patterns across different inputs [Smilkov et al., 2017]. This enhancement makes it easier to identify features that consistently affect the model's decisions, thus improving the explainability and reliability of the results.

1.4 How XAI Can Bridge the Alignment Gap

The diversity of computational XAI techniques underscores the multifaceted nature of model interpretation, with each method providing unique insights into the model's behavior. Adding to that the fact that there are many ANN models of the human ventral stream, we are left with the question: Which model/explainable artificial intelligence method combination is best aligned with explaining the strategies used by the ventral stream of primates in object recognition tasks?

This study is dedicated to a detailed examination of differences within explanations generated by ANNs. Our objective is twofold: initially, to evaluate the degree of variance among ANN explanations by comparing them against one another, and to identify the most appropriate metric that encapsulates the differences between such explanations. This endeavor is challenged by the inherent dissimilarity of explanation methods, which are not inherently designed for comparative analysis. Subsequently, our research endeavors to ascertain the optimal combination of XAI methodologies and ANNs that most accurately emulate the cognitive strategies employed by primates in visual recognition tasks. This dual-faceted approach not only contributes to the understanding of the explainability of ANNs but also bridges a crucial gap in the modelling and approximating biological cognitive processes by artificial systems.

In this project, we aim to elucidate the extent of alignment between ANNs and primates in visual recognition tasks. Building on the foundation established by prior studies on behavioral and neural alignment, this work seeks to highlight the specific features within input images that both ANNs and

the human ventral stream rely on for visual recognition. Such an investigation not only promises to enhance our understanding of the interpretative capabilities of XAI methods applied to models but also serves to identify the ANN model that most closely mirrors the functionality of the primate ventral stream in these tasks. The significance of this research lies in its potential to inform both the development of more biologically aligned artificial vision systems and the refinement of computational models for understanding human visual processing.

1.5 Thesis Outline

In this thesis, following the introduction, the structure unfolds as follows:

Chapter 2: Comparative Analysis of Explanation Methods - This chapter presents a represenative examination of different explanation methods employed by artificial neural networks in the context of visual object recognition. The aim is to identify and compare the variations in explanations generated by various ANN models and assess their alignment with human-like visual processing. This comparison is crucial for understanding how different XAI techniques interpret similar visual inputs and how these interpretations can be standardized and improved. The methods section within this chapter introduces the Sharpness Metric, an innovative approach to quantitatively assess the clarity and distinctiveness of different explanations.

Chapter 3: Alignment with Human Vision - In this chapter, the focus shifts to aligning machine vision models with human vision, evaluating how closely the explanations generated by ANNs mimic the visual strategies employed by humans. This chapter delves into the behavioral experiments designed to test the alignment of machine-generated visual strategies with those derived from human cognitive processes. It discusses the development and application of behavioral proxies as a novel method for comparing ANN explanations to human visual strategies without direct access to internal cognitive processes.

Chapter 4: Discussion - This chapter synthesizes the findings from the comparative analysis and alignment studies, discussing the implications of these results for the development of more interpretable and human-like ANNs. It explores the potential applications of these findings in real-world scenarios where AI interacts with humans, such as in autonomous vehicles

and medical imaging, and considers future directions for research in improving the explainability and reliability of machine learning models.

The thesis concludes with a summary of the findings, emphasizing the importance of creating interpretable AI systems that are aligned with human cognitive processes. It highlights the contributions of the study to the fields of cognitive science and artificial intelligence and proposes areas for further investigation.

Chapter 2

Aim 1: To compare the explanations for the decisions made by various ANN models of visual object recognition

2.1 Background

In machine learning models, a variety of methods are available for generating explanations. These methods differ in their approach, and as a consequence, the resulting explanations are distinct from one another. For instance, methods like occlusion and feature ablation work similarly to methods deployed in psychology and behavioral neuroscience, such as the bubbles technique [Gosselin and Schyns, 2001]. They operate by altering a portion of an image and then observing if a neural network can still predict the object in the image to make decisions. Other, more sophisticated methods like integrated gradients [Sundararajan et al., 2017a] operate on individual pixels by generating a path of linear interpolation between a baseline image (usually a black image) and the input image, computing the integral of the gradient of the prediction with respect to the input image along this path. The result is an attribution map that highlights the regions of the input image most crucial for the prediction, with each pixel assigned a score proportional to its contribution. Our primary objective is to understand the variability in model-derived explanations. Specifically, we aim to comprehend the differ-

ences between explanations. This knowledge would enable us to evaluate the explanations generated by multiple model architectures and explanation methods to assess aspects such as consistency, uniqueness of explanations, and specificity of explanation to an image. The crux of this challenge lies in the establishment of a robust framework capable of quantifying the disparities among various explanation techniques. By achieving this, it becomes feasible to systematically evaluate and, consequently, select the most efficacious methods for interpreting the decision-making processes of ANNs. Such an endeavor is paramount for advancing the explainability and trustworthiness of machine learning models, thereby enhancing their applicability in diverse domains.

The necessity to devise a comparative framework for explanation methodologies emerges from the critical need to refine the interpretive processes applied to machine learning models. When practitioners encounter limitations or deficiencies in their current explanation techniques, the ability to differentiate between similar methodologies and pivot to more effective, divergent approaches becomes invaluable. This transition, however, is hindered by the fundamental challenge that existing explanation methods were not conceptualized with interoperability and direct comparability as core design principles.

This lack of a pre-established standard for preprocessing explanation outputs for comparative analysis leads to significant challenges in ensuring equitable evaluations. Without a clear guideline on how to normalize or preprocess these explanations, drawing meaningful comparisons becomes a complex task fraught with potential inaccuracies and biases.

Moreover, the question of which specific features of the explanations should be the focus of comparison further complicates the issue. Identifying the most salient aspects of explanations that meaningfully contribute to their interpretative value requires a deep understanding of both the technical underpinnings of the methods and the practical contexts in which they are applied.

Finally, the selection of appropriate methodologies for comparing these salient features is another critical piece of the puzzle. The choice of comparison metrics and statistical methods must be carefully considered to ensure that the comparisons are not only valid but also relevant to the stakeholders involved.

In this study, we explore the efficacy of different families of similarity measures in assessing the disparities among explanations generated by various machine learning interpretative methods. Our investigation is driven by the

hypothesis that significant variations exist in the performance of these families, thereby influencing the choice of optimal evaluative metrics for specific contexts within the realm of machine learning explainability.

To this end, our analysis will be anchored in two principal families of similarity measures: statistical and perceptual. Within the statistical domain, we will employ measures such as the L2 Norm and L1 Norm, which facilitate a direct comparison of pixel values in the generated explanations. These measures are grounded in classical statistical analysis, offering a straightforward, quantifiable approach to similarity assessment.

Conversely, our exploration of perceptual similarity measures will include techniques such as the Learned Perceptual Image Patch Similarity (LPIPS). This approach leverages neural networks to abstract high-level features from the underlying image data, thereby enabling a comparison that transcends mere pixel value differences to encapsulate more nuanced, perceptually relevant distinctions.

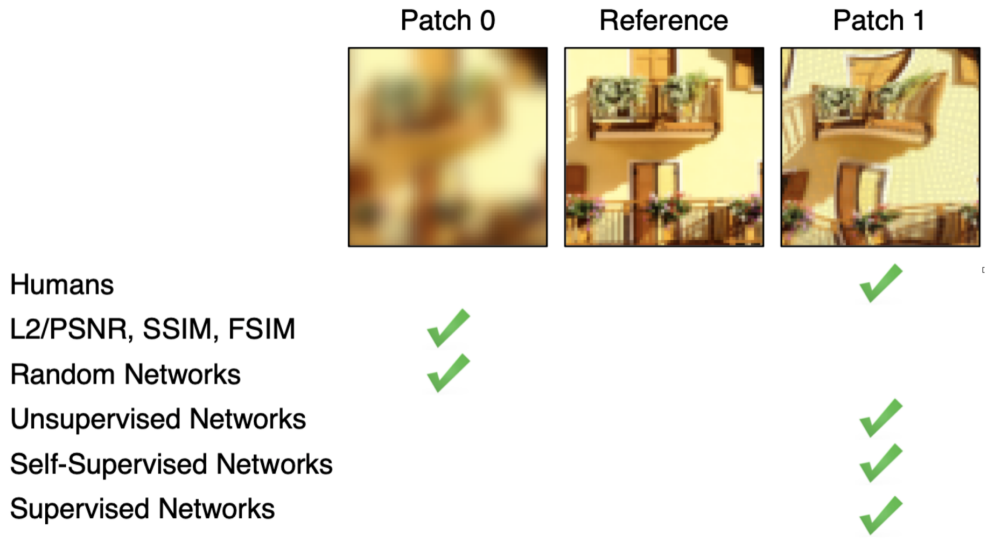

Figure 2.1: Comparison of image patch similarity judgments using different assessment methods. Patch 0 and Patch 1 are compared to a Reference image to determine similarity. Humans, as well as different network methodologies including L2/PSNR [Horé and Ziou, 2010], SSIM [Horé and Ziou, 2013], FSIM [Zhang et al., 2011], Random Networks [LeCun et al., 1998], Unsupervised Networks [Hinton and Salakhutdinov, 2006], Self-Supervised Networks [Misra and van der Maaten, 2020], and Supervised Networks [Krizhevsky et al., 2012b], are used for assessment. The checks indicate which patch the distance method predicts is more similar to the reference images. Deep neural network methods are aligned with human judgment of distance. Figure adapted from [Zhang et al., 2018]

Figure 2.1 [Zhang et al., 2018] provides a representative evaluation of different image patch similarity assessment methods by comparing their outcomes against human judgments. The image showcases two patches, Patch 0 and Patch 1, each compared to a Reference image to determine their similarity. This comparison is assessed across several methodologies including traditional statistical measures like L2/PSNR \citep{hore2010image}, SSIM \citep{hore2013relationship}, FSIM \citep{zhang2011fsim}, and more modern neural network approaches such as Random Networks \citep{lecun1998gradient}, Unsupervised Networks \citep{hinton2006reducing}, Self-Supervised Networks

\citep{misra2020self}, and Supervised Networks \citep{krizhevsky2012imagenet}.

The checks in the table indicate whether each method aligns with human judgments in recognizing similarities between the patches and the reference image. Notably, only trained deep learning based methods consistently align with human judgments across both Patch 0 and Patch 1, suggesting that these methods may offer more reliable and human-like interpretations of visual similarity.

This figure illustrates the efficacy of various computational approaches in mirroring human perceptual assessments, highlighting the potential strengths of self-supervised and supervised learning techniques in tasks that require nuanced visual understanding. Such insights are crucial for developing AI systems that better emulate human cognitive processes, particularly in fields requiring precise image analysis and interpretation.

By juxtaposing these families of similarity measures, our work aims to shed light on their relative merits and limitations in the context of evaluating machine learning model explanations.

2.2 Hypothesis

This inquiry is anchored in the premise that discernible differences exist in the effectiveness of different similarity measure families used to evaluate the variances between explanations generated by various methods. Such an examination is crucial for revealing the complex aspects of model explanation assessments and for pinpointing the most fitting families of similarity measures for specific application scenarios.

H0 (Null Hypothesis): There is no significant difference in the effectiveness of similarity measure families in evaluating the discrepancies between explanations from a range of machine learning model explanation methodologies.

H1 (Alternative Hypothesis): Significant differences are evident in the effectiveness of similarity measure families for evaluating the variances between machine learning model explanations generated by diverse methods, with at least one family of similarity measures demonstrating superior efficacy in providing clear and actionable evaluations compared to others.

This hypothesis lays the foundation for an in-depth evaluative framework aimed at measuring and comparing the effectiveness of various families of similarity measures in identifying distinctions between explanations from

multiple methodologies. By methodically analyzing these differences, the research seeks to advance the field of machine learning explainability, facilitating more precise and informed selections of similarity measure families, and thereby enhancing the clarity and trust in model explanations.

2.3 Methods

We employed a suite of 10 XAI measures, as delineated in the introductory section, to generate interpretative analyses for a cohort of 200 images sourced from the COCO image dataset. These images were selected to encompass a diverse array of 10 object categories, specifically: bear, elephant, person, car, dog, apple, chair, plane, bird, and zebra. All the analysis was done on explanations from the ResNet-50 architecture.

In our preliminary examination, we conducted a direct comparison of explanations by leveraging both families of image similarity measures. The aim was to ascertain the relative effectiveness of these methodologies in differentiating between explanatory outputs. To support this evaluation, we devised a new metric named "Sharpness", conceptualized as the proportion of instances in our analysis matrix where an explanation, upon comparison with its corresponding explanation, manifested a lower average distance than all other comparative instances within the same row.

2.3.1 Introduction to Sharpness Metric

Sharpness is defined as:

$$\text{Sharpness} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(d(E_{iA}, E_{iB}) < \frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N} d(E_{iA}, E_{jB}) \right)$$

where:

- N is the total number of images in the dataset,
- E_i represents the explanation corresponding to the i^{th} image,
- $d(E_i, E_j)$ denotes the distance measure between the explanations for images i and j,

- $\mathbf{1}(\cdot)$ is the indicator function, yielding 1 if the condition within is true, and 0 otherwise.

This 'Sharpness' metric thereby encapsulates the precision with which each methodology discerns between explanations, emphasizing those instances where an image's explanations are most distinctly identified from their counterparts (Figure 2.2).

2.3.2 Challenges to measuring distance

Our investigation revealed that both statistical and perceptual distance measures lacked precision. Specifically, when applied to two images generated by the same explanation, the distance was unexpectedly high. In spite of the fact that in visualizations, those explanations looked like they hi lighted the same areas to human observation.

This phenomenon can be attributed primarily to two factors:

Disparate Numeric Scales: Each explanatory method operates within its unique numeric range. For instance, one method might generate explanation tensors with values spanning from -1 to 1, whereas another method might produce tensors with values ranging from 0 to 250. Distribution of Emphasized Image Regions: Methods based on gradients assign values to every pixel in an image, including minimal values to less significant pixels. This granularity, albeit minor, significantly impacts the efficacy of some distance metrics.

2.3.3 Normalization and Focus on Salient Pixels

To address these challenges, we implemented two corrective measures:

Normalization: We standardized the explanation tensors across all methods, scaling them to a uniform range between 0 and 1. Focus on Salient Pixels: We introduced Explanation Masked Images (EMIs) to preserve only the most critical pixels from the original image, enhancing the relevance and clarity of the explanations.

2.3.4 Explanation Masked Images (EMIs) and Their Generation

To generate EMIs, we first generated filtered versions of the original images by only retaining the top percentiles (50, 60, 90, etc.) of the highly informative pixels (given the feature attribution map of each explanation) for ResNet-50 (inspired from earlier work by Hooker et al., 2019). We refer to them as **explanation masked images** (EMI). The creation of EMIs involves a two-step process. The first step *explanation generation*, involves applying standard explanation methods, such as Saliency [Simonyan et al., 2013] or Occlusion [Zeiler and Fergus, 2014], to generate an explanation for a given image. These explanations provide a means to rank pixels based on their significance in the image's overall context. In the second step, *percentile cutoff calculation and separation*, we determine percentile cutoffs to segregate pixels into different significance tiers. For instance, a 95th percentile cutoff would separate the top 5% of pixels (in terms of importance, as per the explanation) from the rest. Using these cutoffs, we generate two distinct types of EMIs for each image: **Positive EMI, pEMI**: This image encompasses the top 'x' percentile of pixels deemed significant by the explanation. It retains the features considered most crucial in the original image. **Negative EMI, nEMI:** Conversely, this image includes the lower '100 - x' percentile of pixels, highlighting the features deemed less important. To ensure robustness and versatility in our approach, we experiment with a range of cutoffs. This variety allows us to thoroughly evaluate the impact of significant versus non-significant features in image explanations (Figure 2.3).

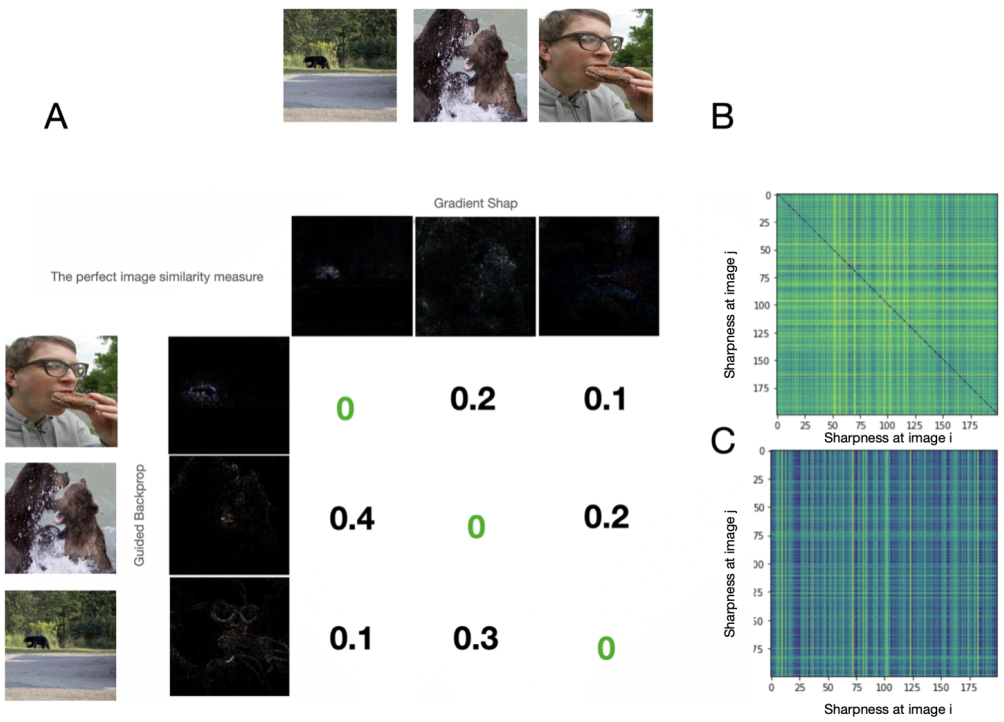

Figure 2.2: This figure demonstrates the evaluation of similarity between visual explanations derived from different explainability methods for images. Panel A displays a schematic similarity matrix with high sharpness where a value of '0' along the diagonal indicates that the visual explanations from two different methods for the same image are most similar to each other, relative to comparisons with other images. Panels B and C depict heatmaps representing the sharpness scores of raw explanations and EMIs, respectively. These scores assess the clarity and distinctiveness of the visual explanations provided by each method. Comparing EMIs is more conducive to high sharpness scores that comparing raw explanations.

2.3.5 Addressing Discrepancies in Explanatory Methodologies

Discrepancies in explanation methodologies, wherein one might designate the most critical pixels with a red hue and another with blue, could poten-

tially skew distance metrics. Using EMIs circumvented this wherein the most salient pixels were substituted with their counterparts from the original images. This approach ensured uniformity, such that a pixel at position [0, 0] highlighted by one XAI technique retained identical value when accentuated by another.

The determination of an appropriate threshold for saliency demarcation presented a subsequent challenge. This was addressed by aggregating the explanations generated by a given explainability method to procure percentile rankings for each pixel. Rather than presupposing an optimal cutoff, a strategy of sampling various thresholds was adopted, acknowledging the speculative nature of selecting the most suitable demarcation.

2.3.6 Enhanced Analysis Using EMIs

Using EMIs yielded markedly enhanced sharpness over direct evaluation of the raw explanations. Leveraging this methodology to facilitate comparison of explanations, the study was expanded to encompass a diverse array of distance metrics. Within the statistical category, the following metrics were employed: L1 Norm, L2 Norm, Structural Similarity Index Measure (SSIM), Root Mean Square Error (RMSE), and Peak Signal-to-Noise Ratio (PSNR). In the realm of perceptual metrics, the investigation was limited to the application of Squeezenet-based LIPIPS, as the utilization of AlexNet and the VGG-16network was precluded due to their roles in generating explanations for subsequent experiments.

For each conceivable pairwise comparison among the explanations of the ten XAI measures, a distinct matrix was generated, culminating in a total of 45 comparison matrices for each similarity measures. With six similarity measures in consideration, this approach resulted in the production of 270 matrices. This computation was replicated across 15 EMI thresholds, thus each threshold yielded the same number of matrices. Eeach matrix facilitated the derivation of a singular sharpness score, embodying the comparative clarity or distinctiveness brought about by the respective similarity measure.

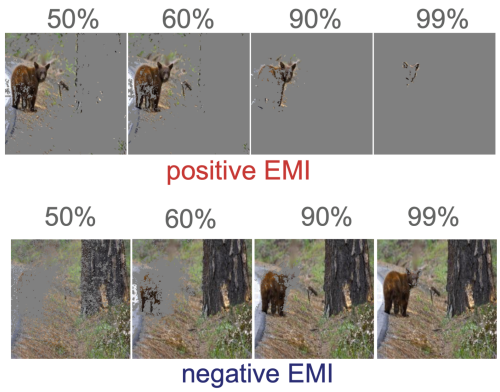

Figure 2.3: A schematic explaining how the appearance of EMIs changes with the percentile cutoff for both positive and negative EMIs

2.3.7 Methodological Framework for Comparative Analysis

The methodology employed for the comparative analysis of ANN explanations can be succinctly delineated into the following sequential steps: Initially, explanations for the dataset images were generated employing the entire spectrum of available XAI methodologies. Subsequently, these explanation heatmaps were transformed into EMIs. The third step involved selecting explanations derived from a pair of explanation methods, and for each image, comparing its EMI against those of all other images generated by the alternate explanation method, employing a singular image similarity measure for each comparison. The fourth step entailed the computation of sharpness for each resultant matrix. Lastly, these sharpness scores served as a basis to allow the high level analysis necessary to test the hypothesis.

2.3.8 Comparative Analysis of Raw Explanations and EMIs

To determine the more efficacious approach between comparing raw explanations and EMIs, an analytical strategy was employed wherein 270 sharpness scores were computed for pairwise comparisons using raw explanations, and an equivalent set of 270 sharpness scores was generated for EMIs, specifically at the postivie EMIs of the 50th percentile cutoff. Following the computa-

tion, the analysis moved to the visualization stage, where a histogram was plotted to encapsulate and contrast the sharpness scores derived from both methodologies. This histogram served as a graphical representation, facilitating an intuitive comparison of the distribution and magnitude of sharpness scores associated with raw explanations versus those obtained through EMIs, thereby guiding the decision-making process regarding the optimal comparison framework.

2.3.9 Evaluating Similarity Measure Families

To address the inquiry regarding the selection of the most suitable family of similarity measures, an approach was taken wherein the sharpness scores attributed to the statistical measures were averaged. This consolidated metric was then juxtaposed against the sharpness scores obtained from the LPIPS measure across all EMI cutoffs. This comparative analysis was visualized through plotting, which facilitated a direct comparison between the average sharpness of the statistical measures and the perceptual sharpness as represented by LPIPS at each EMI threshold. The anticipation underlying this approach was that the superior family of similarity measures would manifest as notably sharper in these plots, thereby providing a clear visual indication of its efficacy in comparison to its counterpart.

2.4 Results

The results, as visualized in figure 2.4, are outlined below:

Distribution of Explanation Sharpness The quantitative assessment of explanation sharpness was visualized in histograms and analyzed through statistical measures. As depicted in Figure 2.4, two distinct patterns emerged when comparing raw explanations to explanations modified by EMIs at the 50th percentile cutoff.

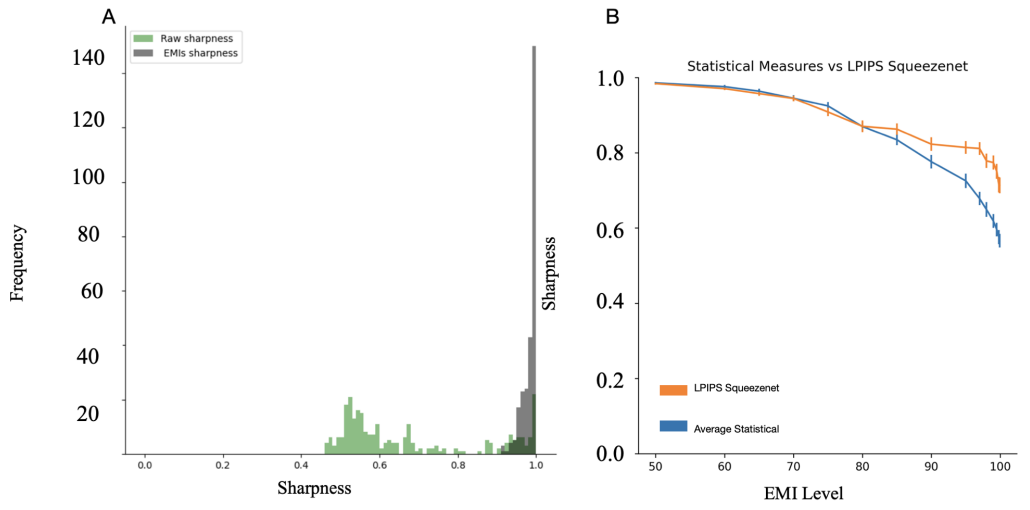

Figure 2.4: A: Histogram of sharpness score distributions comparing raw explanations and EMIs. The frequency of sharpness scores is plotted on the y-axis, with the sharpness score itself on the x-axis. The green bars represent the distribution of sharpness scores for raw explanation comparisons, while the grey bars denote the scores for EMIs at the 50th percentile cutoff. The stark contrast between the two methodologies is evidenced by the dense aggregation of EMI sharpness scores towards the higher end of the scale, signifying a greater degree of discriminability relative to the raw explanation scores. B Sharpness scores comparison between statistical measures and LPIPS Squeezenet across different EMI levels. The y-axis denotes the sharpness scores, while the EMI levels are plotted along the x-axis. The average sharpness scores for statistical measures are shown by the blue line, and the LPIPS Squeezenet scores by the orange line. The error bars represent the standard error of the mean, illustrating the precision of the sharpness score estimates at each EMI level. The declining trend in sharpness with increasing EMI levels is evident for both measures, with statistical measures consistently outperforming LPIPS Squeezenet until LPIPS overtakes it at the higher EMI levels

Raw Explanations: The frequency distribution of sharpness scores for raw explanations is characterized by a broad range with a moderate average sharpness score of 0.679 ± 0.181. This indicates a significant variability in the

sharpness and discriminability of the explanations, suggesting inconsistencies in their quality and clarity.

EMI Explanations: In stark contrast, the EMI explanations displayed a highly concentrated distribution near the maximum sharpness score, with a mean value of 0.986 ± 0.018. This near-maximal sharpness and minimal variance underscore a consistent delivery of high-fidelity explanations, affirming the efficacy of EMIs in enhancing the sharpness and reliability of neural network interpretations.

Sharpness Trends Across EMI Levels The evaluation of sharpness across varying EMI levels further elucidated the differential impact of explanation complexity on the explainability of neural network decisions, as shown in Figure 1B.

Decreasing Sharpness with Increasing EMI Levels: A general decremental trend in sharpness was observed as EMI levels increased, with both statistical measures and LPIPS Squeezenet indicating a decline in sharpness from levels 50 to 80. This trend highlights a potential trade-off between explanation complexity and visual clarity.

Pivotal Transition at Higher EMI Levels: A notable transition occurred at the 80th percentile EMI cutoff, where sharpness levels began to increase, surpassing those observed at lower EMI thresholds. This suggests an optimal balance between explanation detail and sharpness, where higher EMIs begin to reveal more distinct and relevant features necessary for clear explanations.

Precision of Measurements: The narrow error bars associated with each data point across the EMI spectrum signify high precision in the sharpness measurements, thereby reinforcing the reliability of the observed trends. The data indicates that the conclusions drawn from these trends are robust and not attributable to random variations.

Comparative Performance of Explanation Methods A comparative analysis of statistical measures and LPIPS Squeezenet revealed a crucial insight into the performance dynamics of different explanation methods across EMI levels. Initially, statistical measures demonstrated superior sharpness; however, as the EMI level increased beyond the 80th percentile, LPIPS Squeezenet began to exhibit a higher sharpness score, overtaking the performance of statistical measures. This crossover highlights the effectiveness of

perceptual metrics like LPIPS Squeezenet in capturing finer details at higher complexities, potentially making it more suitable for applications requiring precise high-level interpretations.

Conclusion

The findings from this study underscore the significant enhancement in explanation sharpness achieved through the use of EMIs, particularly at optimal cutoff levels. The consistency in high sharpness scores across different EMI levels supports the utility of EMIs in providing reliable and precise explanations. This is critical for the explainability and trustworthiness of artificial neural networks, especially in complex domains where clear and accurate explanations are paramount. The study also provides a basis for selecting appropriate explanation methods based on specific needs and complexities, guiding future applications and development in the field of neural network explainability.

2.5 Discussion

The presented results significantly enhance our understanding of the efficacy of various similarity measure families in evaluating explanation discrepancies derived from different machine learning model explanation methodologies. This inquiry revolves around competing hypotheses regarding the performance of these similarity measure families.

Under the null hypothesis (H0), it was postulated that there would be no significant differences in the effectiveness among the similarity measure families. Contrary to this assumption, the analysis revealed compelling trends, particularly at higher EMI cutoffs. Specifically, from the 80th percentile onward, EMIs demonstrated increasingly sharper scores compared to lower cutoffs, indicating a significant variance in the effectiveness of similarity measures used. This pronounced disparity in sharpness at upper EMI cutoffs challenges H0 and suggests that not all similarity measure families perform equivalently. The sharpness of an explanation, a direct measure of its clarity and effectiveness, varied significantly with the type and settings of the similarity measures used, underscoring the inherent differences in how these measures evaluate the nuances of machine learning model explanations.

The alternative hypothesis (H1), which suggests that certain families of

similarity measures demonstrate superior efficacy at specific EMI cutoff levels, finds strong support from our findings. Statistical measures consistently delivered higher sharpness scores compared to LPIPS Squeezenet across a broad range of EMI levels, until a pivotal shift occurred at higher cutoffs. This differential performance not only highlights meaningful distinctions between the similarity measure families but also emphasizes the impact of selecting appropriate EMI levels.

Given these outcomes, the logical course is to reject the null hypothesis in favor of the alternative. The p-value (approximately $6.57\times103016.57\times10301$) underscores the profound influence of similarity measures and their respective EMI cutoffs on the sharpness—and by extension, the clarity and actionability—of evaluations of machine learning model explanations. These findings highlight the necessity of carefully selecting the similarity measure families for evaluating machine learning explanations, particularly when striving for maximum clarity and explainability.

Further, the empirical data has illuminated a strategic pathway for advancing comparative analysis within XAI methodology. The efficacy of utilizing EMIs over raw explanations provides a solid foundation for further experimental endeavors. The discovery that LPIPS, especially at higher EMI cutoffs, offers a sharper and more distinct evaluation than statistical measures sets a new benchmark for our experimental protocols.

This strategic shift toward prioritizing LPIPS in the comparison of EMIs is based on its consistent superiority in achieving sharper evaluations at elevated EMI thresholds. This level of sharpness is crucial for isolating the most relevant features from explanation maps, thereby enhancing the explainability and utility of the insights derived from these analyses. Additionally, this shift encourages a more nuanced exploration of how different similarity measures can be optimized for various types of machine learning explanations, potentially leading to more precise and actionable outcomes.

As we move into the alignment phase of our methodology, we are guided by the insights gained so far. By leveraging the precision of LPIPS and EMIs to compare explanations, we aim to refine our understanding of the intricacies within machine learning explanations and continue to enhance the robustness of comparative analysis in this evolving field. This approach not only aligns with our empirical findings but also positions us to further explore the potential of different similarity measures in bringing about clearer and more actionable machine learning explanations.

Chapter 3

Aim2: Alignment with human vision

3.1 Background

The advent of sophisticated AI systems has brought to the forefront the pressing need for transparency and explainability in their decision-making processes. This necessity stems from the essential role AI plays across various domains, where decisions made by these systems can have significant impacts. XAI emerges as a crucial field, aiming to bridge the gap between AI system operations and human understanding, thereby fostering trust and ensuring the ethical use of AI technologies.

However, various XAI methods yield differing explanations for identical inputs (and tasks) within the same ANN model, as shown in **Figure 1.2**. It is challenging to determine which explanation should be considered as most accurate. To address this, **?** proposed to benchmark the goodness of the machine explanations as the strength of their alignment with human explanations. However, to achieve this, a key assumption is that there are tools to reliably estimate the primate visual system's explanations.

The perceptions of primates are usually measured using psychophysical methods (figure 3.1) such as 'bubbles; and classification images. The bubbles technique [Gosselin and Schyns, 2001] works by presenting the primate subjects with images that have regions randomly obscured by "bubbles." These bubbles effectively mask parts of the image while leaving other parts visible. Over a series of trials, different parts of the images are revealed or concealed,

and the primate's ability to recognize or respond to the image is carefully observed and recorded. By analyzing the areas of the images that were visible during successful recognition or response, researchers can infer which specific features or parts of the image are most critical for the primate's perception. This method allows scientists to gain insights into the visual processing and perceptual mechanisms of primates, shedding light on how they interpret and interact with their environment.

Classification images is a technique that uncovers the visual cues and features primates, including humans, use to make perceptual decisions. In this method, subjects are presented with a series of visual stimuli that contain a mix of signal and noise. The signal represents the actual visual feature being tested, while the noise consists of random visual patterns that do not convey meaningful information. By systematically varying the signal and noise across many trials and analyzing the instances where the subject correctly identifies or responds to the signal, researchers can statistically derive the "classification image." This image effectively highlights the specific visual features and patterns that were most influential in the subject's decision-making process. Through this approach, scientists can gain a deeper understanding of the underlying mechanisms of visual perception and the critical elements that influence how primates, including humans, interpret complex visual scenes.

Prior research indicates that both bubbles [Gosselin and Schyns, 2001] and 'classification images' [Eckstein and Ahumada, 2002], suffer from multiple shortcomings when it comes to estimating human explanations.

The Bubbles technique might miss critical interrelations among visual elements due to its piecemeal exposure of image features. This limitation could lead to an incomplete understanding of object recognition, as it neglects the holistic principles that characterize human perception, where the entirety of an image is often perceived as more than the aggregate of its parts. Such an approach risks overlooking the interconnectedness of image features, potentially skewing the interpretation of their importance.

Conversely, the Classification Images method faces the hurdle of noise artifacts. The method's reliance on aggregating trial data, each with its own variance, can inadvertently enhance noise. This noise accumulation can conceal the actual signal, introducing artifacts in the importance maps that may mislead by suggesting the relevance of non-essential image regions.

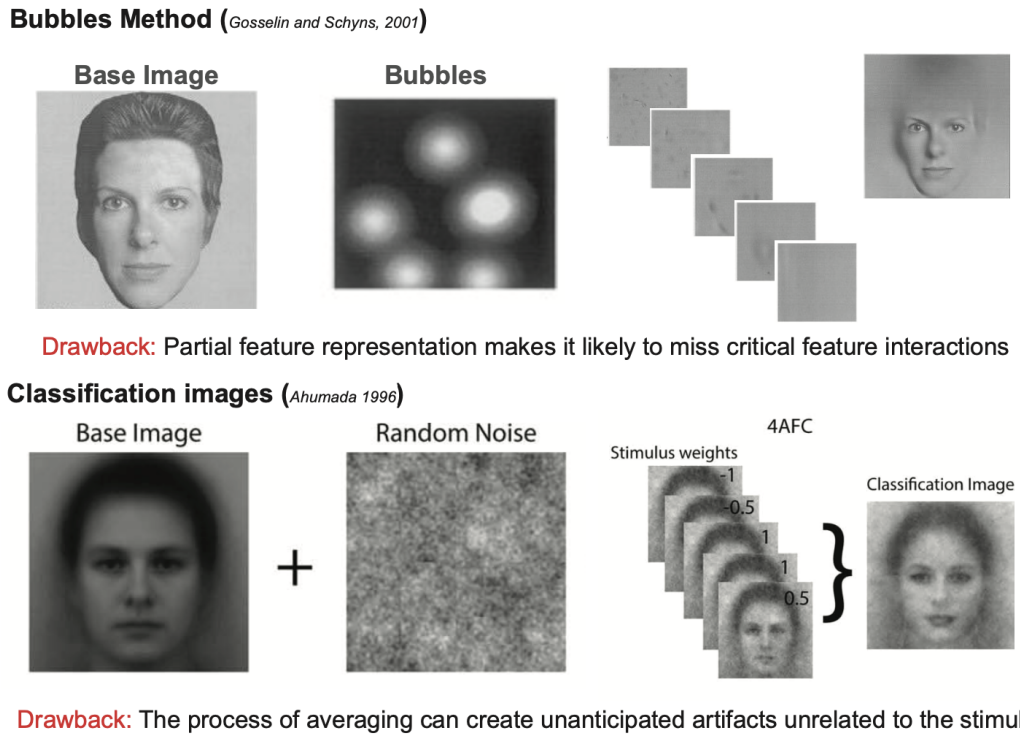

Figure 3.1: The Bubbles Method [Gosselin and Schyns, 2001], shown in the top panel, uses partially occluded versions of an image to reveal which features are most important for recognition tasks. The drawback is that it may overlook critical interactions between features due to the isolated presentation of certain areas. The Classification Images technique [Ahumada, 1996], demonstrated in the bottom panel, combines a base image with random noise and uses human observers to determine the stimulus weights, which are then used to generate a classification image that reveals important features. However, this method has the drawback of potentially introducing artifacts through the averaging process, which may not be related to the actual features important for recognition. In both panels, the image on the right represents the human feature importance map hypothesized by each method. Figure adapted from [Gosselin and Schyns, 2001] and [Ahumada, 1996]

The unreliability of the assumptions underpinning our initial approach posed a significant challenge, necessitating the development of a new method-

ology. This methodology must adeptly capture the visual strategies employed by both the primate visual stream and ANN models designed to emulate the primate visual stream. Achieving such a methodological breakthrough is crucial for creating comparable representations between these two systems. Only with these comparable representations can we facilitate the alignment of explanations, thus giving us new way to benchmark the alignment of ANN models of the ventral stream that will serve as an addition to the neural and behavioral benchmarks that already exist.

We reasoned that if we could use the exact same method to derive an explanation in both the Target (the model that we seek to explain, e.g. say ResNet-50) and the Reference (the model that serves as the gold standard, as we propose – humans) species, then we could avoid the issues mentioned earlier. Given that most XAI tools require full access to the internals of the ANNs, such unrestricted access to human or other primates is not feasible with the current state of the art in psychophysics. Therefore, in this study, we develop a proxy method that aims to bypass this requirement and enable comparing the similarity in explanations between a Target and Reference model without access to its internal components, while specifically focusing on object discrimination behavior.

We aim to develop a method to bypass the direct comparison of explanations across two models (a Target and Reference) with the help of a behavioral strategy. Below, we first explain the method to generate a set of ground truth similarity rankings across XAI generated attribution maps for the two models, and then provide evidence that we can reproduce a significantly similar ranking by comparing the behavioral performance of the models on explanation masked images of varying degrees (obtained only from the Target model).

This would then pave the way for us to compare ANN behavioral proxy to primate behavioral proxy in order to measure alignment of different models and explanation methods with primate vision.

3.1.1 Hypothesis

This projects tries to elucidate the most closely aligned model / XAI method combination for approximating the visual strategies relied on by primates for object recognition tasks.

Null Hypothesis (H0): The behavioral proxies derived from the feature maps of a given ANN explanation method are equally aligned with the under-

lying visual strategies used by humans to detect objects as those derived from other explanation methods. In other words, there is no significant difference in the alignment of visual strategies between the specified ANN explanation method and others.

Alternative Hypothesis (H1): The behavioral proxies derived from the feature maps of a given ANN explanation method are more closely aligned with the underlying visual strategies used by humans to detect objects than those derived from other explanation methods. This suggests a significant difference in alignment, favoring the specified explanation method over others.

We hypothesized that the way such images impact the behavior of two systems might be symptomatic of how similar the underlying explanations used to generate those images are.

3.1.2 Methods

3.1.3 Estimating the true differences in explanations

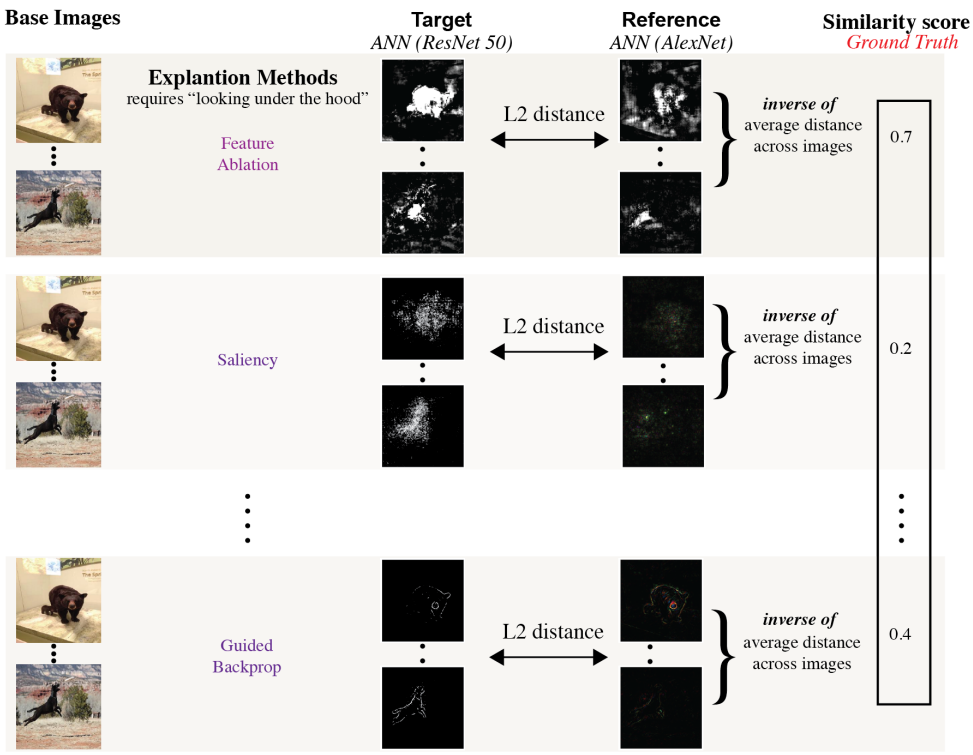

Figure 3.2: **Estimating image similarity between two attribution maps using L2 distance across XAI outputs**. We use the same 200 base images. Here we use ResNet-50 as the Target model and AlexNet as the Reference. For each image, and the object categorization task we estimate the feature attribution maps from the XAI method shown in the column to the right of the base images. We estimate the L2 distance between each of the attribution maps and take the inverse of the mean over all the distances for each XAI tool to assign an overall similarity score. The higher this value, the more similar the explanation outputs are between AlexNet and ResNet-50.

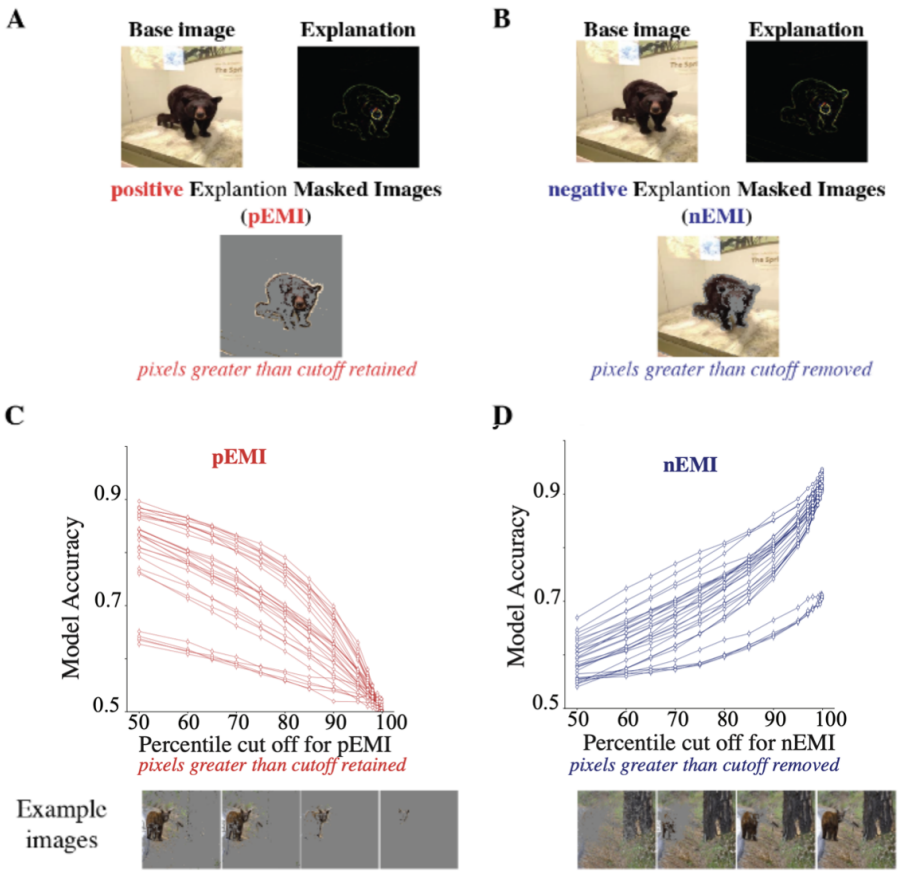

Figure 3.3: **Estimating EMI and validating it with model accuracy tests. A.** Generation of positive (by retaining the pixels greater than a cut off) and negative (by removing the pixels greater than a cut off) EMI. **B.** Model accuracies reflecting the decrease and increase in performance consistent with the expected changes with EMI cut off levels.

We first define a **Target model** (whose explanation we seek, e.g., ResNet-50 [He et al., 2016], **Figure 3.2**) and a **Reference model** (a robust model, whose explanation serves as a gold standard for evaluating the goodness of explanations for the target model). While ultimately, we want to use humans as the reference model, we need an image-computable, fully differentiable model (e.g., AlexNet [Krizhevsky et al., 2012b], **Figure 3.2**) to develop and validate our method. An explanation of a model's output typically takes the form of a heat map that indicates how different features (pixels) of the input

image contribute to the model's output (feature attribution). To test our method, we first estimated the ground truth in how similar the explanations of ResNet-50 (for 200 natural images belonging to 10 object categories, from the MS COCO dataset [Lin et al., 2014], on object discrimination) are to AlexNet, by directly comparing the feature attribution maps produced by ten different explanations (e.g., Saliency, feature ablation, integrated gradients) using a few distance metrics (e.g., L1-norm, L2-norm). This step produces the ground-truth rank order of how explanations compare across ResNet-50 and AlexNet (see right most column in **Figure 3.2**). Once such a ground truth has been established, we now desire a human-compatible procedure ("proxy") to recover the *ground-truth rank order*. A successful proxy method should be able to retrieve this *ground-truth rank order* without "looking under the hood" (i.e., by only probing behavior of the Reference model).

It is worth mentioning that we chose to use L2 norm for the alignment study because it gave good sharpness for the EMI level we were using as mentioned in the last chapter.

3.1.4 Behavior with EMI as proxy

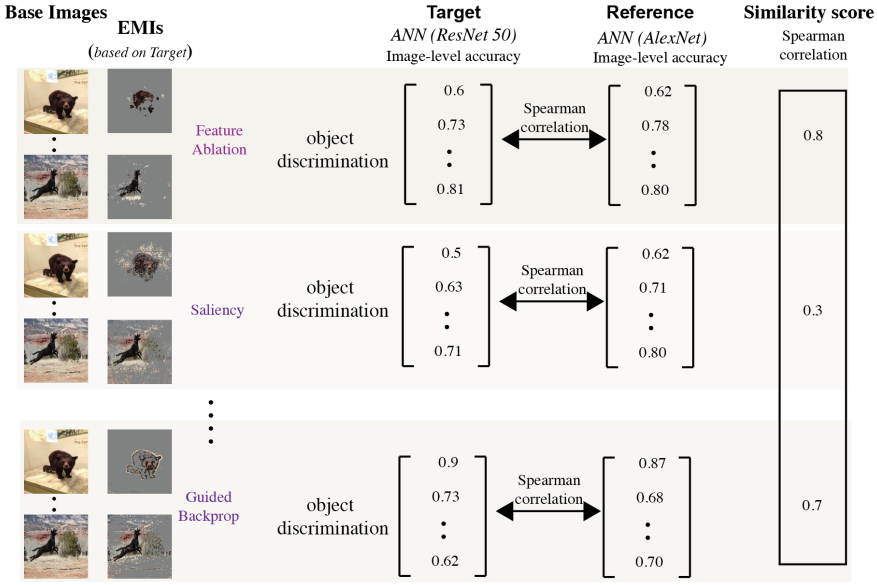

Figure 3.4: **Behavioral tests on EMI.** The EMI generated from each of the explanation methods (for the Target model) are presented to the Target and the Reference model. The image-by-image accuracy pattern is correlated across the models to get a similarity score.

We Approximates the average object discrimination accuracy for that image against all possible distractor objects (referred to as $B.I_1$ in Rajalingham et al., 2018). To do that, our ANN models generate output comprising ten probabilities for each image in a dataset of 200, corresponding to the likelihood of each category within the image set. This probabilistic output is then utilized to compute a behavioral index, henceforth referred to as i1. This computation is critical as it aligns the model's performance metrics with analogous human behavioral assessments.

To further refine our analysis, we introduce a secondary behavioral index, i_2, calculated using the formula:

$$\frac{p(\text{target})}{p(\text{target}) + p(\text{distractor})}$$

Where target refers to the actual category of the image, while distractor

denotes the likelihood assigned to an alternative category. For example, if the target category is 'elephant' and the distractor is 'bear', the behavioral index i_2 for this scenario is computed as the ratio of the probability of 'elephant' to the sum of probabilities of 'elephant' and 'bear'. It is pertinent to note that i_2 is not calculated for the image's ground truth category to maintain distinctness between target and distractor.

Consequently, for each of the 200 images, nine distinct i2 values are calculated, excluding the ground truth category. The primary behavioral index, i1, is then determined by averaging these nine i2 values for each image, thereby encapsulating a single performance metric per image. This metric is significant as it facilitates direct comparison with human subject responses, providing a quantifiable measure of model performance in mimicking human-like behavior in visual categorization tasks.

i1 indices were computed in a manner analogous to the model-based approach, albeit with an initial focus on a binary discrimination task facilitated through the use of the JavaScript library jsPsych and Amazon Mechanical Turk. From the original set of 200 images, 50 were randomly selected for this purpose. This method enabled the generation of i1 indices for human subjects by capturing their decision-making patterns in response to the presented images.

As shown in figure 3.4, this procedure gives us an i1 vector for both our target and distractor models (on the target's EMIs). We correlate the i1 vector at each explanation method using Spearman R correlation to get a similarity score. It is with this similarity score rank order that we are trying to recreate the rank order of direct comparisons.

3.1.5 Choosing an EMI cutoff

Figure 3.5 illustrates the comparison of different neural network architectures using Spearman's rank correlation coefficient to evaluate their ability to recreate the rank order generated from direct comparisons of explanations. The red circles represent pEMI values, while the blue circles represent nEMI values for each architecture. The architectures compared include AlexNet [Krizhevsky et al., 2012b], ResNet-50 [He et al., 2016], VGG-16 [**?**], MobileNet [**?**], SimCLR [Chen et al., 2020], and ViT-b32 [Dosovitskiy et al., 2020]. The results indicate that pEMIs generally perform better at recreating the rank order, as evidenced by higher Spearman R values across all models. This led to the decision to use pEMIs, particularly from the 85th

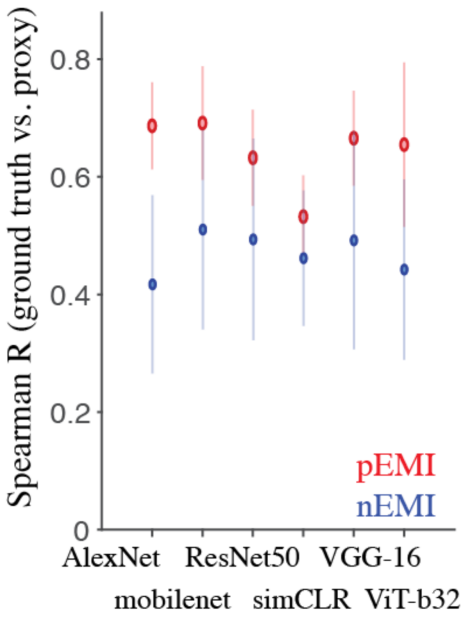

Figure 3.5: Comparison of neural network architectures using Spearman's rank correlation coefficient to recreate the rank order generated from direct comparison of explanations. Generally, using pEMIs is batter at recreating the rank order

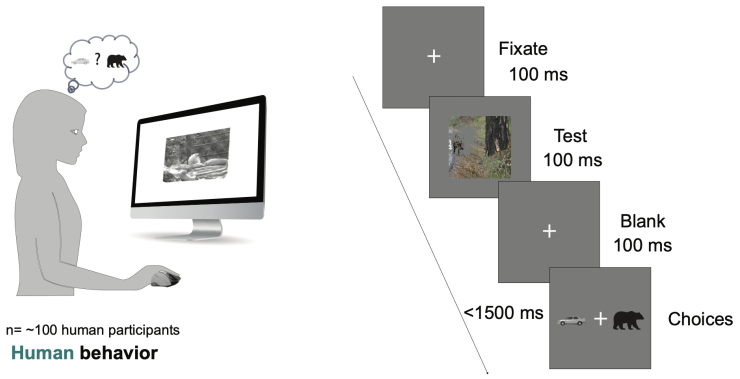

Figure 3.6: This figure shows a human subject undergoing the jsPsych task we used to generate our human accuracy vectors that we used to compare human behavior to machine behavior on target model EMIs

percentile, ensuring the cutoff preserves sufficient context. Therefore, we chose pEMIs because they were better at predicting the rank order from direct comparisons.

Now that the behavioural i1 was established for each cutoff and explanation method across all models. We decided to use EMIs from the 85th percentile. This decision was taken after visually inspecting the EMIs at each cutoff and selecting a cutoff that has enough context preserved.

3.1.6 Measuring Human Behaviour

Now that we have established that our method is architecture agnostic, we decided to treat the human brain as just another reference architecture. To do that, we needed to generate a human accuracy vector for each explanation method that we would correlate with machine accuracy vector for each of our model architectures and all of their explanation methods. Due to cost constraints, we selected only 50 images from our 200 image test set.

50 image indexes were selected at random. this allowed us to select the same 50 EMIs from each of our models and their explanation methods (a total of 7 models x 10 xai methods yielded 70 image sets).

Participants We engaged 100 anonymous participants recruited via the Amazon Mechanical Turk (MTurk) platform, a widely used resource for ob-

taining diverse human cognitive data. All participants were English speakers with normal or corrected-to-normal vision, though specific demographic details such as age and gender were not collected to maintain participant anonymity in accordance with ethical guidelines (Figure 3.6).

Stimuli The stimuli were EMIs generated from the target model's explanations. We made sure to use the same 50 images from all our target model architetures and their explanation methods.

Task Design The task was structured as a series of binary object discrimination challenges. Participants were asked to identify and discriminate between two presented objects. The design involved all possible pairwise combinations of the 50 images, ensuring comprehensive coverage and multiple observations per image. This design aimed to robustly evaluate the discriminative capabilities of both human subjects and computational models under equivalent conditions.

Procedure Each trial began with the display of a fixation point for 500 ms, a standard procedure to stabilize the participant's gaze and reduce initial visual wandering. Following this, a test image was presented centrally for 100 ms, simulating the brief glance conditions typical in everyday visual perception. Immediately after the test image vanished, a choice screen appeared, presenting two images side by side—one the target object and the other a distractor. These images were displayed in canonical views against a neutral background to minimize contextual cues that could influence object recognition. Participants indicated their choice by clicking on the image that they believed corresponded to the previously seen object. The position (left or right) of the correct answer was randomized across trials to prevent positional biases. No feedback was given to participants to prevent learning effects over the course of the experiment.

Data Collection Data was collected in the form of participant selections for each trial, with each participant completing an estimated 350 trials based on the total number of images and tasks. This data collection resulted in approximately 35,000 total trials, providing a substantial dataset for statistical analysis of human visual object recognition performance on target generated EMI stimuli.

Data Analysis For analysis, human accuracy vectors were constructed for each image based on the proportion of correct responses. These vectors represented the human performance baseline for each image and were used for subsequent correlations with machine performance. To that end, the human behavioral single dimensional index was used (i1).

The Behavioral i1 metrics play a critical role in quantifying discriminability at the image level. These metrics were specifically designed to assess the accuracy with which each image is recognized relative to all other objects across the test set, providing a detailed measure of recognition performance for individual images.

Definition and Calculation The Behavioral I1 metric, termed as one-versus-all image-level performance, is computed for each image using a sensitivity index known as `d'` (d-prime) [Kar et al., 2019]. The d' index quantifies an observer's ability to distinguish between signal (correct identification of an image) and noise (incorrect identification as another object). This index is calculated as follows:

d' = Z(Hit Rate) - Z(False Alarm Rate)

Where:

- Z(Hit Rate): The Z-score transformation of the Hit Rate, which is the proportion of trials in which the image was correctly identified as its true object class.
- Z(False Alarm Rate): The Z-score transformation of the False Alarm Rate, which is the proportion of trials in which any other image was incorrectly identified as the object class of the target image.

Data Collection for d' For each of the 50 images, the Hit Rate and False Alarm Rate were computed based on human responses collected during the task. The Hit Rate for each image was directly measured from the trials where the image was presented, and the False Alarm Rate was aggregated from all trials where any image was incorrectly identified as the target object. These rates were then converted into Z-scores using the cumulative Gaussian distribution function, facilitating a standardization that allows for comparisons across different images and tasks.

Application in Analysis The resulting Behavioral i1 scores for each image provide a detailed profile of human discriminability, reflecting the precision with which participants could recognize and distinguish each specific image from all others in the dataset. In the context of this study, these metrics were crucial for correlating human performance with the outputs of computational models. Each AI model's performance was similarly quantified using the Behavioral i1 metric across the same images, enabling a direct comparison to human accuracy. This comparison helps determine how well each model, coupled with its explanation method, aligns with human visual cognitive processes in object recognition tasks.

Ceiling Calculation

The ceiling for human alignment was calculated using a resampling technique to estimate the upper bound of consistency in human discriminability measures, thereby setting a realistic benchmark for AI alignment. Initially, the complete set of human-generated Behavioral i1 scores was randomly split into two subsets multiple times. For each split, a separate Behavioral i1 score was calculated for each subset, resulting in two distinct clusters of i1 values for each random division of the data. These two clusters were then correlated against each other to assess the internal consistency and reliability of human discriminability measures across random samples. This process was repeated extensively, with numerous random splits, generating a series of correlation scores between the paired i1 clusters from each iteration. The mean of these correlation scores was calculated to derive an average correlation coefficient, which serves as an empirical ceiling for human discriminability in our dataset. This ceiling represents the theoretical maximum consistency achievable by humans in object recognition tasks under the experimental conditions, providing a critical reference point for evaluating the alignment of AI models with human visual cognitive processes.

Alternative Methodology

In our methodology, we initially employed grey backgrounds for the EMIs. Upon reflection, it becomes evident that the use of phase-scrambled backgrounds would have been preferable to minimize any residual influences from the silhouettes left by explanations. In order to assess the difference between the two methods of generating EMIs, we reran all of the experiment for our two best performing models (VGG-16 as target and ResNet-50 as reference) However due to the cost constraints, we were unable to rerun the last part, the part where we find out which explanation method from each model is most aligned with human behaviour.

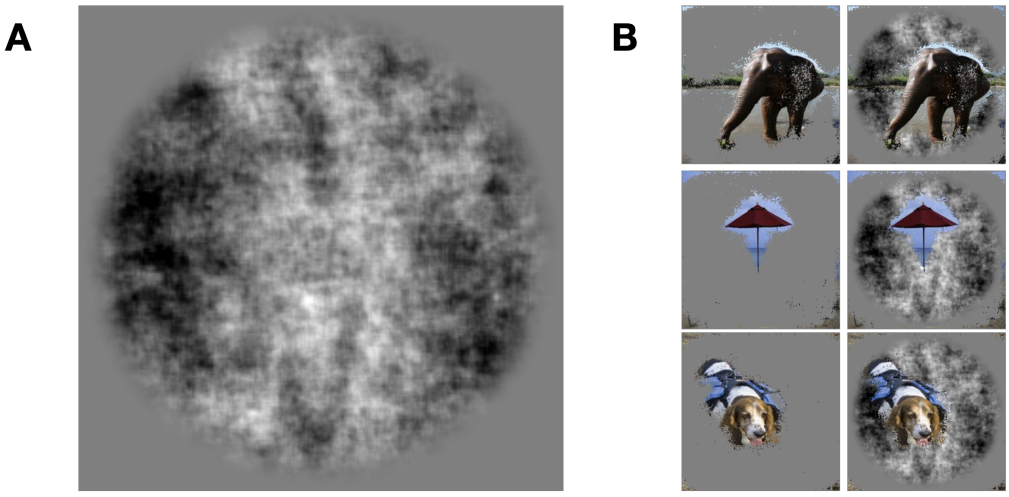

Figure 3.7: This figure shows A: the template we used as a background and B: left column shows 3 example images made using our normal method of making EMIs and on the right column, EMIs made with the phase scrambled background template

The first step was for us to use a template phase scrambled background. This background was used to regenerate EMIs for all of our target models explanations (spanning 10 methods).

After the phase scambled EMIs were generated, we got the behavioral i1 vectors for both target and reference models and correlated them with each other. We then correlated those numbers with the direct comparison distances we computed earlier. This procedure is exactly as described earlier in the methods section of this chapter. this was done for positive EMIs and negative EMIs separately as seen in A and B of figure 3.8.

The error bars, which reflected consistency of behavior with bootstrapping, were smaller. This might indicate that using EMIs with phase scrambled backgrounds are perceived with more consistency. However, looking at C, it is clear that while the error bars were smaller when we used phase scambled images, the general patterns were not changed.

With this in mind and the fact that we are unable to repeat the human data collection due to cost constraints, we can rely on our current mode of

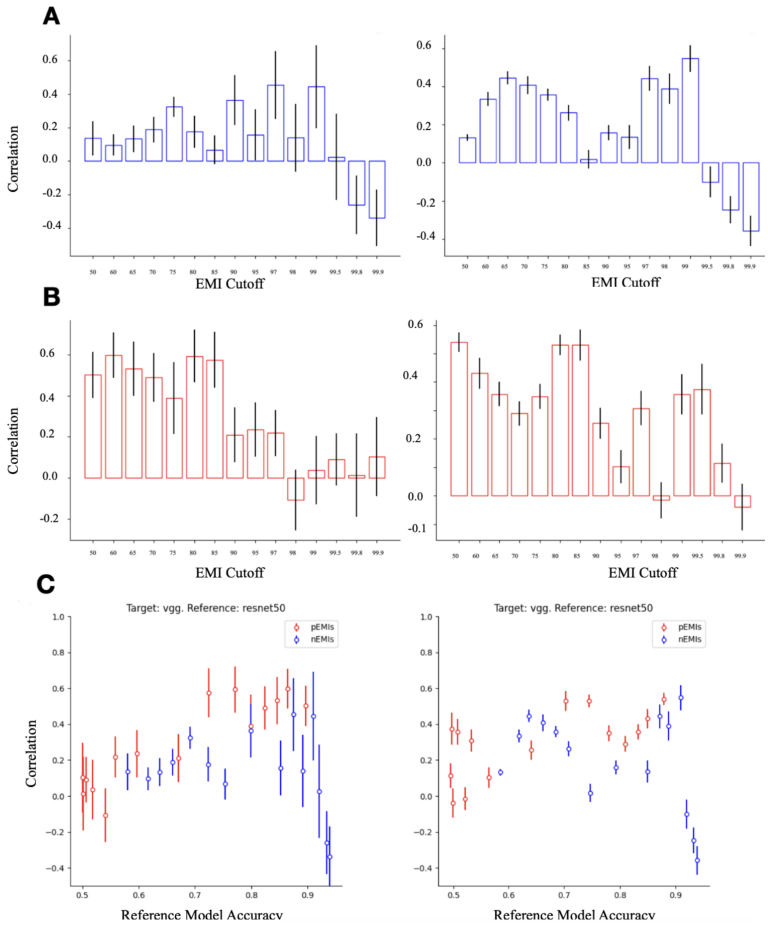

Figure 3.8: The difference in proxy performance between normal EMIs (left) and phase scambled EMIs (right) (A) Blue bar plots showing the correlation between behavioral i1 vectors for target and reference models with direct comparison differences using positive EMIs (pEMIs), across different EMI thresholds represented by numbers on the x-axis. (B) Red bar plots display similar correlations using negative EMIs (nEMIs). Error bars represent consistency measured via bootstrapping, indicating higher perceptual consistency with smaller error bars for phase scrambled images. (C) Scatter plots of Spearman correlation coefficients (R) as a function of reference model accuracy (x-axis) for pEMIs and nEMIs with target model as 'vgg' and reference model as 'ResNet-50', further illustrating that phase scrambling does not alter general behavioral patterns, despite reducing variability in error bars.

generating EMIs while keeping in mind that our results would have been less noisy had we used phase scrambled backgrounds for our EMIs.

3.1.7 Ranking XAI methods

We first isolated the i1s from our ANNs for the same 50 images shown to humans. Spearman correlations were employed to evaluate the alignment between human human i1s and those derived from various AI models, across ten distinct explanation methods. This alignment assessment generated ten correlations for each model, as each method provided a distinct explanatory perspective on the model's decision-making process.

Each set of ten correlations was subsequently ranked from highest to lowest. This ranking aimed to identify which explanation method most closely aligned with human behavior in object recognition for each model. This process was methodically repeated across all seven models employed in the study.

The culmination of this analysis was the identification of the highest ranked correlation for each model. This peak rank pinpointed the explanation method that, among all tested, best mirrored human visual strategies in recognizing objects. This methodical approach allowed for a systematic comparison of explanation methods across models, highlighting the relative effectiveness of each method in approximating human-like interpretative processes through measuring behavior.

3.2 Results

We validated our surrogate approach by comparing explanations from a Target model to various Reference models. The comparison involved calculating the L2 distance metrics across feature attribution maps from different XAI methods to establish a ground truth rank order of similarity. Subsequently, we correlated this with the behavioral accuracies of ANNs tested on the generated EMIs using Spearman correlation. The results showed a significant correlation (Spearman $R \approx 0.7$ across all models), confirming that our method effectively captured the similarity of explanations across models (Figure 3.10).

In a parallel empirical test, we evaluated human subjects' object discrimination performances using EMIs derived from different ANN models and XAI

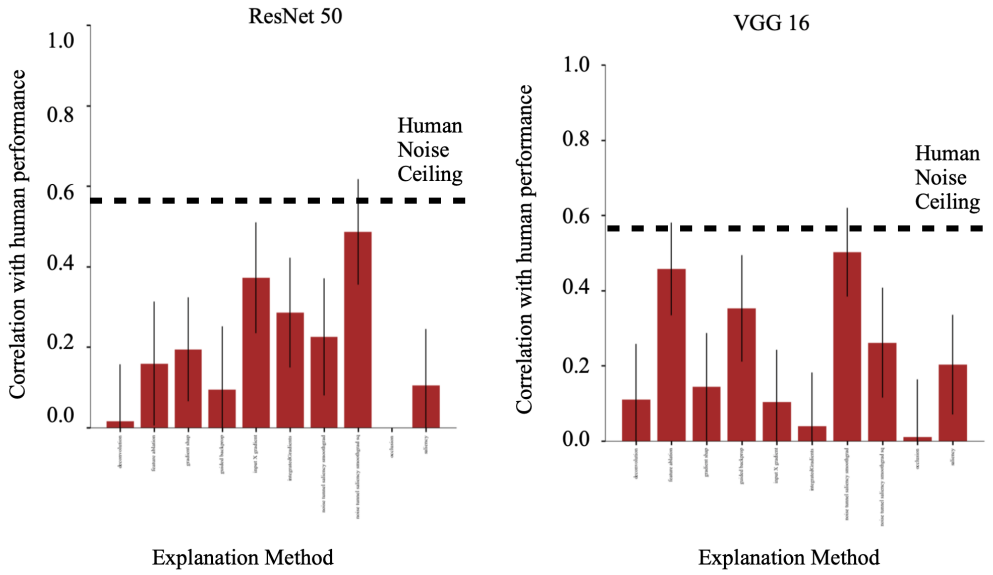

Figure 3.9: This figure compares the alignment of two deep learning models, ResNet-50 and VGG-16, with human behavioral data across various explanation methods. Each bar in the histograms represents the consistency of a specific method with human behavior, quantified on a scale from 0 to 1, where 1 indicates perfect alignment. The "Human Noise Ceiling" represented by the dashed line indicates the highest performance expected due to inherent variability in human judgments, providing a benchmark for comparing model performance. Error bars reflect the variability of results, obtained through bootstrapping, showcasing the stability and reliability of each method's alignment with human behaviors.

methods. The configuration employing VGG-16 with the Saliency-method (noise tunneling smooth gradient) yielded the highest alignment with human behavioral patterns, achieving a Spearman correlation of 0.45. This outcome empirically identified the most closely human-aligned model among those tested, based on its performance on pEMIs.

3.3 Discussion

In the present study, we have successfully demonstrated that artificial neural networks (ANNs), when paired with appropriate XAI methods can serve as effective proxies for dissecting human visual strategies during object recognition tasks. Our approach, which centers around the creation and use of positive Explanation Masked Images (pEMIs), illuminates key aspects of human visual processing by capturing the most informative pixels as determined by various XAI methods.

The choice to focus on pEMIs was driven by our goal to understand the critical features that ANNs rely on, which we hypothesized would closely align with human perceptual processes. Notably, our best-performing model approached the noise ceiling of human behavior, highlighting its potential to closely mimic human visual recognition. This significant correlation observed between the behavioral accuracies of ANNs on pEMIs and the ground truth rank order of similarity suggests that the features deemed important by these models are indeed reflective of those used in human vision. This finding supports the use of pEMIs as a robust tool in bridging the gap between human and machine vision, particularly in contexts where direct comparison between human and ANN strategies is challenging.

Furthermore, the distinct performance of the VGG-16 model under the Saliency-method (noise tunneling smooth gradient) in aligning with human behaviors underscores the potential of specific XAI methods in enhancing our understanding of how ANNs can model human-like decision-making. This result is particularly compelling as it suggests that not all XAI methods are equally capable of revealing the decision-making processes in a manner that is consistent with human cognitive processes. The superior performance of the noise tunneling smooth gradient method may be indicative of its ability to better approximate the nuanced visual cues that humans use to recognize objects.

The alignment of the VGG-16 model with human visual processing strate-

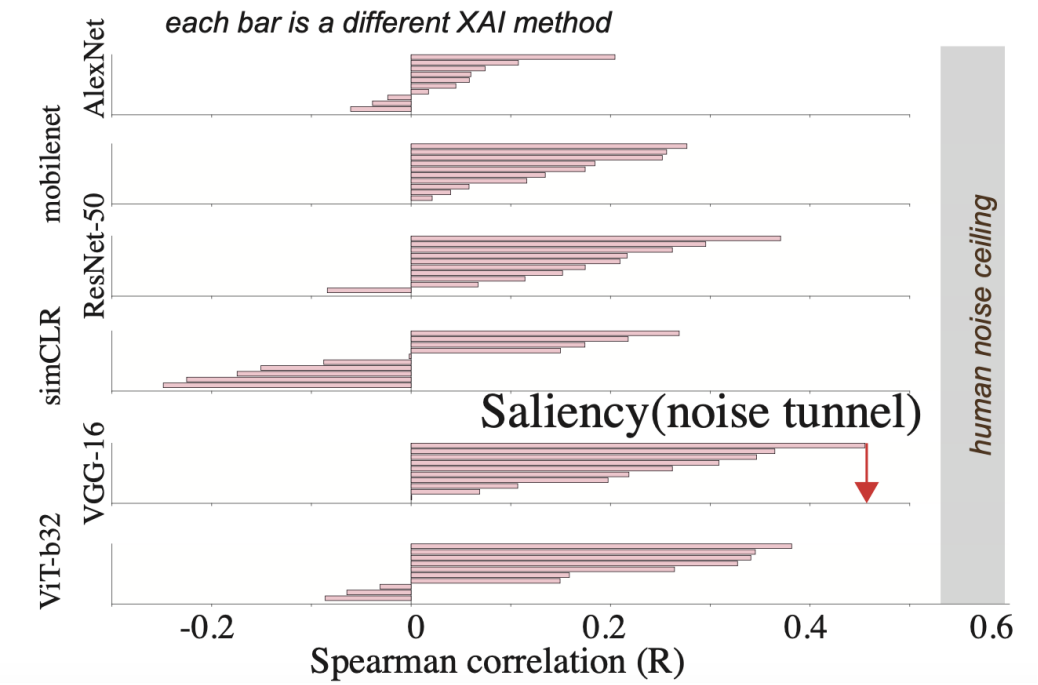

Figure 3.10: This figure represents the Spearman correlation coefficients (R) illustrating the alignment between various artificial neural network models and human behavioral patterns. Each bar corresponds to a different ANN model, with the length representing the strength of the correlation between the model's predictions and human object recognition behaviors. The results are derived from testing with Explanation Masked Images (EMIs) generated by the Salience method using noise tunneling. The arrow indicates the model with the highest correlation, suggesting its superior ability to mimic human visual strategies as assessed in this study. Notably, the VGG-16 model exhibits the most significant alignment with human behavior, consistent with the findings presented in the paper, which indicate it as the model most closely approximating human visual processing among those tested.

gies approached the empirically derived ceiling of human discriminability. This proximity raises intriguing possibilities for both the potential of ANNs to mimic human cognition and the methods we employ to measure and understand this alignment. Given that VGG-16 nearly reached the ceiling, it suggests that our model is effective in emulating human visual recognition patterns under the test conditions. However, this also highlights an opportunity to refine our understanding of the upper limits of human performance. Expanding our dataset and increasing the robustness of our human discriminability measurements could potentially raise the ceiling, providing a more stringent benchmark for future models. Additionally, to further narrow the gap between our best-performing model and the human ceiling, we propose using neurally aligned ANN models of the ventral stream. By enhancing both our measurement ceiling and the human-likeness of our models, we aim to push the boundaries of what artificial systems can achieve in terms of closely replicating human cognitive capabilities in complex visual tasks. It is also worth pointing out that, as seen in figure 3.11, highly aligned explanations look like they keep a lot more human understandable features.

The implications of these findings extend beyond the technical achievements. They offer a methodological framework for future studies aiming to compare and contrast the object recognition strategies of biological and artificial systems. This framework could be particularly useful for refining ANNs to make them not only perform better but also operate in a manner that is more interpretable and analogous to human cognition. Given the proximity of our best model to the human noise ceiling, there is a promising avenue to close this gap even further by employing models that are more aligned with human cognitive processes.

Finally, while our results are promising, they also highlight the inherent complexity in fully modelling human visual strategies. Future research should explore the integration of more diverse XAI methods and ANN architectures to expand the robustness and applicability of the findings. Additionally, incorporating more granular behavioral metrics and diversifying the datasets used could help in further refining the models to capture an even broader spectrum of human visual strategies.

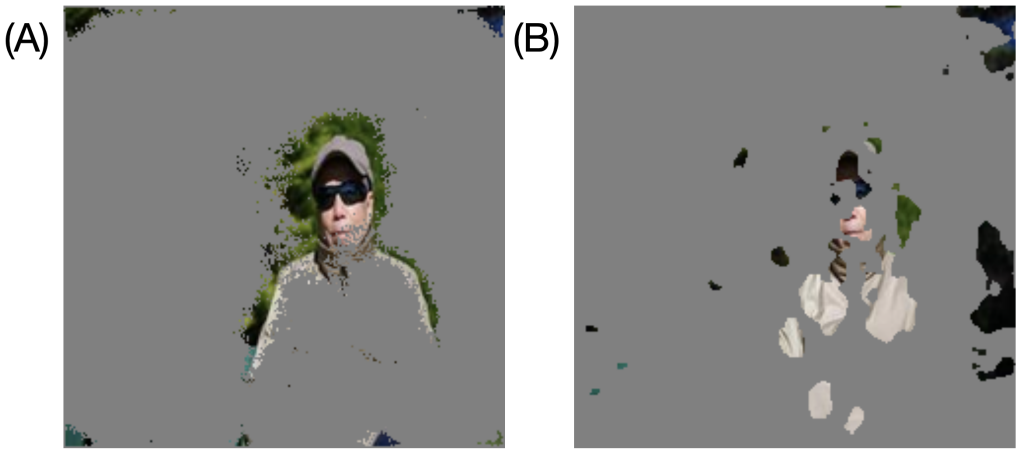

Figure 3.11: The figure illustrates two distinct XAI methods on the same image of a person via the VGG16 model. Panel A displays the noise tunnel smooth grad saliency method, which effectively retains and highlights features crucial for human object recognition. Panel B, in contrast, utilizes the occlusion method, which is less effective in preserving these important features. The comparison underscores the superiority of the approach shown in Panel A for aligning with human visual recognition processes.

Chapter 4

Discussion

4.1 Our Aims in Perspective

This research endeavor was designed with the primary objective of elucidating two fundamental areas of inquiry. Initially, the study sought to explore the methodologies employed in the comparative analysis of explanations derived from ANN models, specifically those designed to emulate the functionalities of the human ventral stream. This stream plays a pivotal role in the primate visual system, primarily in object recognition, and understanding its computational analogs offers profound insights into both artificial and biological information processing systems. The secondary goal of this investigation was to meticulously assess the alignment between various explanation-generation methodologies employed by ANNs and the intrinsic visual strategies employed by primates during object recognition tasks.

The journey towards achieving these objectives was faced with unforeseen technical challenges that necessitated the development and implementation of novel methodological approaches. The initial hypothesis posited that the feature attribution maps, which serve as a cornerstone for the explanations generated from the employed ANN models, would be inherently comparable. However, this assumption was quickly challenged upon encountering significant disparities in the data, which were not amenable to direct comparison. This obstacle led to a rigorous exploration of various normalization techniques, ultimately culminating in the adoption of explanation masked images as a viable solution. This approach enabled a more effective comparison of the feature attribution maps, thereby facilitating a deeper understanding of

the explanations generated by ANN models.

The second aim of the study was met with a similarly daunting challenge. The original premise was predicated on the notion that explanations generated by human subjects could serve as a valid comparative framework for those produced by ANN models. This assumption was, however, met with substantial skepticism upon closer examination. The inherent problems with psychophysical techniques of generating human explanations rendered the direct comparison with ANN-generated explanations problematic. To navigate this impediment, we devised a proxy method, which allowed for an indirect yet robust comparison between the two modalities of explanation generation to shed more light on the visual strategies relied on by the ventral stream and its ANN models.

This work lays the groundwork for future investigations into the intricate relationship between artificial and biological visual systems, with the potential to unveil novel insights into the mechanisms underlying object recognition and cognitive processing. More than that, we posit this this kind of alignment serves as a third check of ventral stream ANN model quality beyond the two established methods of comparing them behaviorally and to neural activations.

4.2 Comparing Explanations

The findings from our research significantly contribute to our understanding of how different groups of similarity measures impact the assessment of discrepancies in explanations generated by various machine learning model explanation techniques.

Initially, it was assumed under the null hypothesis (H0) that no significant variance would be observed in the performance of different similarity measure families. However, the data reveals a notable trend where Explanation Map Indices (EMIs) at higher percentile thresholds, particularly above the 80th percentile, yield more distinct scores compared to lower thresholds. This trend highlights a noticeable difference in the precision of explanations, which directly correlates to their effectiveness. The marked contrast in precision at higher EMI thresholds challenges the initial assumption that all similarity measure families are equally effective.

The observations lend support to the alternative hypothesis (H1), suggesting that certain groups of similarity measures, especially at specific EMI

thresholds, are more effective. The statistical measures are on par with LPIPS Squeezenet in terms of precision across most EMI levels. However, LPIPS Squeezenet outperforms statstical methods at high EMI cutoffs. This variance in performance underscores the existence of differences in how similarity measure families assess machine learning model explanations.

Based on these observations, it appears reasonable to dismiss the null hypothesis in favor of the alternative. The results imply that the selection of similarity measures and their application at various EMI thresholds can greatly influence the precision of the evaluations, thereby affecting the clarity and usefulness of machine learning model explanations. These findings highlight the importance of carefully choosing similarity measure families for evaluating machine learning explanations, especially when clarity and explainability are crucial.

Our empirical analysis has paved the way for a more refined approach to comparing explanations within machine learning models. The proven effectiveness of using EMIs, instead of raw explanations, sets a solid foundation for future experiments. Given that LPIPS, particularly at higher EMI thresholds, provides a clearer distinction than statistical measures, it will be adopted as the benchmark for comparing EMIs in future studies.

This strategic shift towards favoring LPIPS for comparing EMIs is based on its consistently superior performance in producing precise evaluations at higher thresholds. Such precision is key to identifying the most relevant features in explanation maps, thereby improving the explainability and applicability of the insights gained.

As we move into the next phase of alignment, our approach will be guided by the insights obtained so far. By leveraging the precision of LPIPS at high EMI thresholds, we aim to deepen our understanding of the nuances in machine learning explanations and enhance the effectiveness of comparative analyses in this field.

4.3 Alignment with Human Perception

In the context of understanding human visual strategies via explainability methods for Artificial Neural Networks, our research has led to several crucial insights and methodologies that refine how we compare and interpret explanations generated by different XAI methods. This comprehensive exploration begins with the acknowledgment that comparing explanations di-

rectly is inherently complex. The nuanced nature of visual strategies and the variability in how different XAI techniques elucidate these strategies render straightforward comparisons ineffective.

To address this complexity, we introduced Explanation Masked Images (EMIs) as a novel tool to mediate the comparison of explanations from various XAI methods. EMIs represent a strategic distillation of visual data, preserving only the most critical features identified by XAI methods. This approach allows for a more focused and meaningful comparison of how different models perceive and process visual information.

A significant challenge in this line of research has been the 'human explanation gap'—the divergence between human intuitive visual processing and the machine-generated explanations. To bridge this gap, we developed a proxy method that does not require 'looking under the hood' of the model, thus maintaining a realistic scenario for application. This proxy was instrumental in generating a rank order of model explanations by their alignment with human perception, which was crucial for our analyses.

Through rigorous testing, we established that the combination of a specific XAI method and ANN model—specifically, the VGG-16 model using a Saliency-based method (with noise tunneling smooth gradient)—was most closely aligned with human visual processing strategies. This alignment suggests that certain architectural features of VGG-16, combined with the specificity of the saliency method, are particularly conducive to mimicking human-like visual interpretation patterns.

Speculating on the reasons for this alignment, it is plausible that the architectural characteristics of VGG-16, which include deep layering and a robust feature extraction process (Also found in ResNet-50, our second most aligned model), resonate more closely with the hierarchical processing of visual information in the human brain. This hypothesis aligns with emerging evidence suggesting that certain deep learning architectures can parallel human neural activity, particularly in tasks involving complex visual processing.

The implications of this alignment extend beyond theoretical interest; they challenge and complement existing paradigms in both behavioral and neural alignment methodologies. Traditionally, behavioral methods focus on the outcomes of visual processing (e.g., object recognition accuracy), while neural methods often seek correlations between neural activations and model layer activations. Our approach, leveraging both behavioral outcomes through EMIs and structural similarities via architectural alignment, offers a hybrid strategy that may provide a more holistic understanding of how

artificial systems can emulate human vision.

This dual approach not only bridges the gap between computational models and biological vision but also offers a methodological framework that could be adapted to explore other cognitive behaviors. By integrating insights from both behavioral and architectural perspectives, we can enhance the explainability and applicability of machine learning models in tasks that require a nuanced understanding of human-like processing. This alignment, therefore, not only validates our methodological innovations but also sets a precedent for future explorations into the convergence of artificial intelligence and human cognition.

While our study provides significant insights into aligning artificial neural network models with human visual processing strategies, it is important to acknowledge the inherent limitations of these models. First, despite our efforts to mimic human visual perception, the simplification required to model such complex processes means that perfect alignment with human vision is unattainable. Additionally, the models we employ, including VGG-16, operate under constraints that do not exist in human visual processing, such as fixed architecture and predetermined processing pathways, which may not fully capture the dynamic and adaptive nature of biological vision systems. Furthermore, our reliance on proxy methods and EMIs, although effective, introduces a level of abstraction that may omit subtle yet critical aspects of human visual interpretation. These limitations highlight the need for continuous refinement of both the models and the methodologies employed, emphasizing that these findings should be viewed as approximations rather than exact replications of human cognitive processes.

4.4 Future Directions

1. **Testing Additional Human-Aligned Models:** Future research should extend the evaluation framework to include a broader range of ANN models, particularly those already suspected to be closely aligned with human cognitive processes. This could involve testing lesser-known or emerging architectures that have shown promise in preliminary studies. The goal would be to verify and quantify their alignment using the methodologies developed in this research, such as EMIs and proxy comparison techniques.

2. **Integration of More XAI Methods:** Expanding the range of Explainable Artificial Intelligence methods tested can provide deeper insights into how different approaches capture or reflect human-like processing. Including methods that offer different perspectives on decision-making processes within ANNs, such as counterfactual explanations and feature interaction methods, could enrich the understanding of model behavior in complex visual tasks.

3. **Behavioral and Neural Correlation Studies:** Further studies could integrate more comprehensive behavioral and neural correlation analyses to assess the alignment between ANNs and human visual processing. This would involve more intricate experimental setups where human neural activity is measured in response to the same stimuli processed by the ANNs, providing a direct comparison of activity patterns.

4. **Development of Standardized Testing Frameworks:** Establishing a standardized testing framework for evaluating the human alignment of ANN models could facilitate more consistent and comparable research in this field. Such a framework could include standardized datasets, evaluation metrics, and benchmarks that specifically measure human-like processing accuracy and efficiency.

By pursuing these future directions, we can continue to refine our understanding of the parallels between artificial and biological cognitive systems, thereby advancing both the development of more sophisticated AI models and our understanding of human cognition. This alignment holds the promise of developing AI systems that are not only effective but also intuitive and interpretable, bridging the gap between human and machine learning.

Bibliography

Fairness and machine learning: Limitations and opportunities. http://fairnessml.org, 2019.

Apple vision pro. Apple Website, 2024. URL https://www.apple.com. Product overview of the Apple Vision Pro, featuring advanced virtual reality capabilities. Accessed: 2024-07-24.

Saranya A. and Subhashini R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230, June 2023. doi: 10.1016/j.dajour.2023.100230. URL https://doi.org/10.1016/j.dajour.2023.100230. Accessed: 2024-07-24.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020.

A. J. Ahumada. Perceptual classification images from vernier acuity masked by noise. *Perception*, 25(1_suppl):2–2, 1996. doi: 10.1068/v96l0501. URL https://doi.org/10.1068/v96l0501.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence, April 2021. URL https://ec.europa.eu/info/policies/justice-and-fundamental-rights/computers-and-digital-technology/artificial-intelligence/strategy_en. The first ever legal framework on AI proposed by the

European Commission aiming to categorize AI risks into four levels and ensure trust and global competitiveness. Accessed: 2024-07-24.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Miguel P Eckstein and Albert J Ahumada. Classification images: A tool to analyze visual strategies. *Journal of vision*, 2(1):i–i, 2002.

Frédéric Gosselin and Philippe G Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17):2261–2271, 2001.

Isha Hameed, Samuel Sharpe, Daniel Barcklow, Justin Au-Yeung, Sahil Verma, Jocelyn Huang, Brian Barr, and C. Bayan Bruss. Based-xai: Breaking ablation studies down for explainable artificial intelligence, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

Alain Horé and Djemel Ziou. Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure? *IET Image Processing*, 7(12):12–24, 2013.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Kohitij Kar and James J. DiCarlo. The quest for an integrated set of neural mechanisms underlying object recognition in primates. *Annual Review of Vision Science*, 10:19.1–19.31, 2024. doi: 10.1146/annurev-vision-112823-030616. URL https://doi.org/10.1146/annurev-vision-112823-030616.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6): 974–983, 2019.

S. M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, 2024. URL https://arxiv.org/abs/2202.01602.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012a.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012b.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fer-

gus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

Richard F. Murray. Classification images: A review. *Journal of Vision*, 11 (5):2–2, 2011.

Richard F. Murray. Classification images and bubbles images in the generalized linear model. *Journal of Vision*, 12(7):2, Oct 2017. doi: 10.1167/12.7.2. Accessed: Month Day, Year.

Richard F. Murray and Jason M. Gold. Troubles with bubbles. *Vision Research*, 44(5):461–470, 2004. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2003.10.006.

Ian E. Nielsen, Dimah Dera, Ghulam Rasool, Nidhal Bouaynaya, and Ravi P. Ramachandran. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, 2022. doi: 10.1109/MSP.2022.3142719.

C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

Martin Schrimpf and Jon Prescott-Roy. Brain-score. `https://www.brainscore.com`.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017.

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

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

JD Smith. Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin Review*, 21(2):312–331, Apr 2014. doi: 10.3758/s13423-013-0506-0.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328, 2017a.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017b.

Max Wertheimer. Experimental studies of the perception of movement. *Zeitschrift für Psychologie*, 61:161–265, 1912. Original title in German: Experimentelle Studien über das Sehen von Bewegung.

D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.