Perceptual Parallels: Investigating Human-like Phenomena in Deep Neural Networks for Object Recognition

Katha Rohan Reddy Aryan Bansal Venika Sruthi Annam

***Abstract*—Deep neural networks (DNNs) have greatly advanced computer vision, showing similarities in object representations to visual cortical areas in the brain. However, it remains unclear if these representations capture qualitative patterns akin to human perception or brain representations. We explore this by analyzing well-known perceptual and neural phenomena through distance comparisons in both randomly initialized and object recognition-trained DNNs. Our findings reveal that while certain phenomena emerge in both randomly initialized and trained networks, others are exclusively present in trained networks. Moreover, we investigate various DNN architectures to assess their ability to replicate human-like perception. These results provide insights into the conditions necessary for the emergence of perceptual phenomena in both brains and DNNs, highlighting avenues for improving DNNs by incorporating human-like properties.**

I. Introduction

Convolutional or deep neural networks have transformed computer vision, exhibiting impressive accuracy in object recognition tasks and aligning roughly with certain aspects of brain function. However, they still fall short of human performance and display consistent deviations from human perception at finer levels. These differences may stem from either insufficient training data, limited constraints, or architectural shortcomings, prompting exploration into potential qualitative disparities in how visual information is represented between brains and deep networks.

To address this, we leverage classic findings from visual psychology and neuroscience, reframing them as tests of distances between images within underlying representations. For example, the Thatcher effect—a face appearing grotesque when its parts are inverted—can be assessed by measuring the distance between normal and Thatcherized faces. By applying these tests to both randomly initialized and trained deep networks, we aim to discern which properties arise from architecture versus training.

Our investigation spans various perceptual and neural phenomena, categorized into groups such as object statistics, neuron tuning properties, feature relations, 3D shape processing, and global structure. Using a state-of-the-art pre-trained convolutional network optimized for object classification (VGG-16) and its randomly initialized counterpart, we analyze the presence of these properties across network layers. Notably, some properties emerge in intermediate layers but not in the final classification layer, hinting at their importance for intermediate computations.

By systematically examining these properties across different network architectures and distance metrics, we offer insights into the interplay between architecture, training, and emergent properties in deep networks vis-à-vis human perception.

II. Experiments

A. Weber's Law

1) Explanation of Weber's Law and Model Selection: Weber's law, a fundamental principle in perceptual psychology, posits that the sensitivity to changes in a sensory stimulus is proportional to the baseline level of that stimulus. In the context of visual perception, Weber's law suggests that the ability to detect differences in attributes such as length or intensity depends on the magnitude of those differences relative to the initial baseline levels.

For our investigation, we chose to examine the adherence of deep neural networks (DNNs) to Weber's law, specifically focusing on the VGG-16, ViT (Vision Transformer), ResNet (Residual Network), and Inception models. The VGG-16 network, with its deep architecture and widespread usage, served as our primary model of interest. We also included ViT, ResNet, and Inception to explore potential variations across different neural network architectures.

Model	Score
ViT	-0.00527972
ResNet	-0.06
Inception	0.15692122

2) Analysis of Results: Our analysis revealed intriguing insights into the behavior of deep neural networks with respect to Weber's law for length changes in visual stimuli. Initially, we observed a negative correlation difference between pairwise distances and absolute changes in length in the early layers of the VGG-16 network. This indicated a greater sensitivity to absolute changes rather than relative changes in length.

However, as we progressed to the later layers of the VGG-16 network, we found a reversal in this trend. The correlation difference aligned more closely with Weber's law, suggesting that these layers became increasingly sensitive to relative changes in length. Remarkably, this phenomenon was absent in randomly initialized versions of the VGG-16 network, highlighting the crucial role of training in shaping the network's adherence to Weber's law.

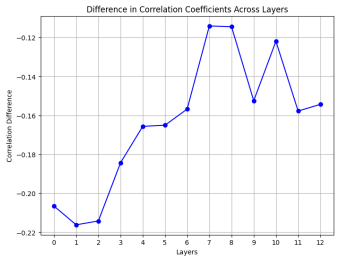

Fig. 1. Weber's law

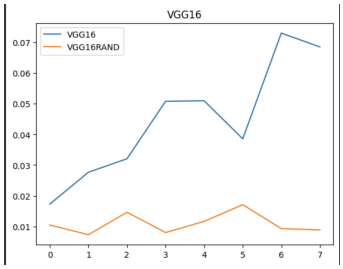

Fig. 2. Thacther distances

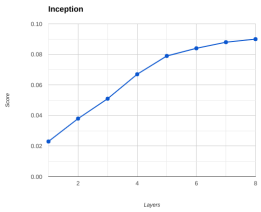

Fig. 3. Inception

In contrast, our analysis of other DNN architectures revealed varying degrees of adherence to Weber's law for length changes. ViT exhibited a slightly negative score, indicating a deviation from Weber's law. ResNet showed a more pronounced deviation with a negative score, while Inception demonstrated a positive score, suggesting a closer alignment with Weber's law.

B. Screen Incongruence

Methodology:

To investigate the sensitivity of deep neural networks (DNNs) to scene context, we utilized images previously tested on human participants. These images featured objects placed against congruent and incongruent backgrounds, aiming to mimic real-world scenarios where objects are encountered within specific contexts. The stimuli comprised 40 objects, with 17 sourced from the Davenport study and 23 from the Munneke study. Objects lacking matching category labels in the ImageNet database were excluded, resulting in the inclusion of 29 objects for analysis.

For each object, two versions were created: one with a congruent background and another with an incongruent background. The congruent background aligned contextually with the object, while the incongruent background presented a context mismatch. This setup aimed to simulate scenarios where an object is encountered in a fitting or mismatched environment.

We employed the VGG-16 network to assess classification accuracy for objects in congruent and incongruent scenes. Specifically, we utilized the final layer (Layer 38) of the VGG-16 network, which returns probability scores for all 1000 categories in the ImageNet database. We measured top-1 accuracy, representing the average accuracy of the ground-truth object label matching the object class with the highest probability. Additionally, we calculated top-5 accuracy, indicating the average accuracy of the ground-truth object appearing among the top 5 object classes with the highest probability.

Human accuracy data, obtained from previous studies, were also analyzed for comparison. Human participants were tasked with naming the objects presented in congruent and incongruent scenes.

Results:

Our analysis revealed notable differences in classification accuracy between congruent and incongruent scenes for both the VGG-16 network and human participants. The VGG-16 network exhibited a substantial drop in accuracy for objects presented in incongruent scenes compared to congruent scenes. Specifically, we observed a 27

Similarly, human accuracy also decreased for objects in incongruent scenes, albeit to a lesser extent compared to the VGG-16 network. Human participants experienced a 14-13

Moreover, our analysis of scene incongruence effects revealed that incongruent scenes were further away from the average feature vector for the object compared to congruent scenes. This effect was particularly pronounced in later layers of all networks, indicating a progressive integration of scene context during processing. Overall, our findings suggest that DNNs, like humans, are sensitive to scene incongruence, albeit to a greater degree, highlighting their susceptibility to contextual influences.

C. Tatcher Effect

The "tatcher effect" in cognitive science, also known as the "expertise reversal effect," refers to a phenomenon where instructional methods or learning strategies that are effective for novices might become less effective or even detrimental for learners with more expertise or prior knowledge in a particular domain. This effect challenges the assumption that one instructional method fits all learners regardless of their level of expertise. In this project we aimed to see if we could repeat the same with DNNs as well

Clearly there is an effect on number of layers against the distance it shows that as layers increase so do the variance betweeen upward and inverted images with the formula as

$$\frac{(d_{\text{up}} - d_{\text{inv}})}{(d_{\text{up}} + d_{\text{inv}})}$$

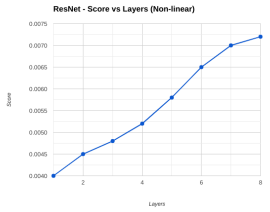

Fig. 4. Resnet

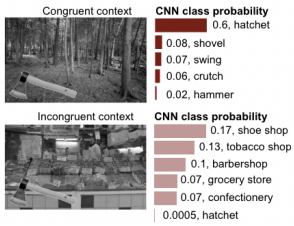

Fig. 5. Screen incongruence

D. Multiple object normalization

Multiple Object Normalization (MON) is a technique in computer vision designed to tackle the complexities of detecting and recognizing multiple objects within images. It operates on the principle of contextual understanding, recognizing that objects often exist within intricate visual contexts where they interact with other objects. By normalizing the representation of each object relative to its surroundings or neighboring objects, MON enhances the network's contextual comprehension of the scene. Additionally, MON considers the spatial relationships between objects, including factors like proximity, occlusion, and relative size, to accurately differentiate and localize individual objects within the image. Moreover, MON addresses variations in object scale, appearance, and orientation by normalizing object features across different instances and configurations, thereby enhancing the network's robustness to such variations. By incorporating these contextual normalization techniques into the network architecture, MON facilitates more efficient and effective learning of object representations, leading to improved performance in various tasks such as object detection, recognition, and segmentation, particularly in scenarios involving multiple objects with complex spatial relationships.

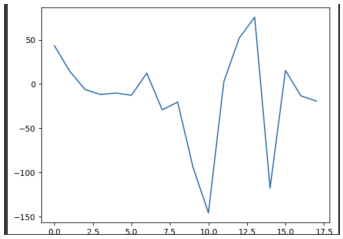

Fig. 6. Features deviation from combined figure

E. Mirror Confusion

The "Mirror Confusion" refers to a phenomenon where vertical mirror images are perceived as more similar than horizontal ones, observed in human perception.

This experiment aimed to investigate mirror confusion in deep neural networks (DNN's), drawing parallels with human perception and neural responses observed in monkeys. In the original paper, the author's utilized the VGG-16 network architecture and a randomly initialized network. Stimuli comprised of images of objects, including naturalistic objects and their rotated versions. Each object was paired with its horizontal and vertical mirror images to assess mirror confusion.

To quantify mirror confusion, a mirror confusion index was defined as:

$$\frac{(d_{\text{horizontal}} - d_{\text{vertical}})}{(d_{\text{horizontal}} + d_{\text{vertical}})}$$

where $d_{\text{horizontal}}$ and d_{vertical} represent the distances between object and horizontal mirror image, and object and vertical mirror image, respectively.

This index was calculated across network layers. Results indicated a positive mirror confusion index, suggesting stronger mirror confusion for vertical mirror images compared to horizontal ones, mirroring human perception. Notably, this trend was absent in the randomly initialized network. The methodology aligned with previous studies on mirror confusion in monkey inferior temporal neurons, validating the approach's relevance in studying neural representations of visual stimuli. (Fig. 7)

We expanded upon the study by employing the identical dataset utilized by the original authors and broadened the analysis to encompass additional models such as ResNet, Inception, and ViT. Our investigation involved analyzing the confusion index values to discern the degree to which these models emulate human perception. We instantiated the experiment using the PyTorch framework, in contrast to the authors' utilization of MATLAB for implementation.

1) VGG16: Our initial model selection was VGG16, characterized by its deep architecture comprising 16 convolutional and fully connected layers. Subsequently, we fed the original, horizontally mirrored, and vertically mirrored images into the pretrained VGG16 model. Extracting the outputs from each layer, we employed the formula stipulated in the original paper to compute the mirror confusion index for each image at every layer. These indices were then graphically represented through plots. (Fig. 8)

Subsequently, we computed the outputs of the final layers and determined the mirror confusion index values. These values were then visualized using a violin chart (Fig. 9), providing a comprehensive representation of the mirror confusion across the final layers of the model.

Mean: 0.279
Std: 0.167
Min: -0.241
Max: 0.673

2) ResNet: For the second model, ResNet was employed, which is renowned for its deep architecture featuring residual connections. ResNet comprises multiple residual blocks, each containing convolutional layers and shortcut connections.

During experimentation, the original image, along with horizontally and vertically mirrored counterparts, was inputted into the ResNet model. Subsequently, outputs from the final layer were extracted, and the confusion index was computed utilizing the prescribed methodology.

To offer a comprehensive depiction of the mirror confusion, the distribution of confusion index values was visualized using a violin plot (Fig. 10), allowing for a comparative analysis of the model's performance across different mirror image orientations. The distribution was pretty similar to that of VGG16.
Mean: 0.263
Std: 0.143
Min: -0.105
Max: 0.625

3) Inception: For the third model, Inception architecture, specifically the PyTorch implementation, was utilized. Inception networks are characterized by their utilization of multiple parallel convolutional pathways, allowing for the extraction of features at different scales.

Similar to the ResNet implementation, the experiment involved feeding the original image alongside horizontally and vertically mirrored versions into the Inception model. Subsequently, outputs from the final layer were extracted, and the confusion index was computed using the established methodology.

To provide a comparative analysis, the distribution of confusion index values was visualized using a violin plot, mirroring the methodology employed with the ResNet model (Fig. 11). This approach facilitated an evaluation of the Inception model's performance in capturing mirror confusion relative to the other models. The inception model showed more perceptual similarity to human brains than ResNet and VGG16.
Mean: 0.384
Std: 0.177
Min: -0.113
Max: 0.827

4) ViT: The last model employed was Vision Transformer (ViT) from the timm library. ViT is distinctive for its attention mechanism, which enables it to process images as sequences of patches rather than using convolutional layers. This architecture allows ViT to capture long-range dependencies and global context efficiently, making it particularly effective for image classification tasks, especially with large datasets.

The experimental procedure remained consistent with previous models: the original image, along with horizontally and vertically mirrored versions, was inputted into the ViT model. Outputs from the final layer were extracted, and the confusion index was calculated. The resulting distribution of confusion index values was then visualized using a violin plot for comparative analysis (Fig. 12). The ViT model showed the most perceptual similarity to human brains out of all the models we tested.
Mean: 0.531
Std: 0.192
Min: -0.126
Max: 0.886

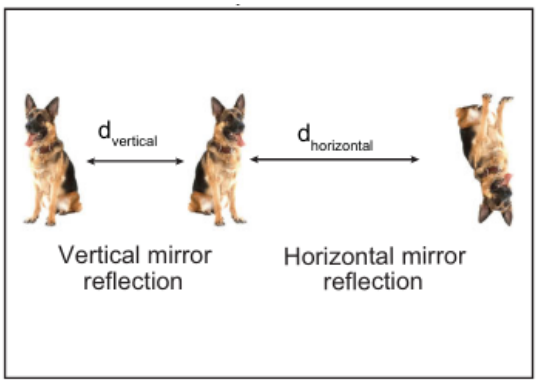

Fig. 7. Mirror Confusion

Conclusion: In summary, our investigation into mirror perception, which examines how the human brain processes visual stimuli, revealed compelling parallels between the behaviors exhibited by the Deep Neural Networks (DNNs) under scrutiny and human perceptual tendencies. This experiment delved into the phenomenon where vertical mirror images are often perceived as more akin to their originals compared to horizontal mirror images, mirroring findings observed in human psychology and neurobiology.

Throughout our analysis, we introduced the concept of the confusion index as a quantitative measure to assess the degree of similarity between original images and their respective mirrored counterparts. By applying this index across various DNN architectures—specifically, VGG16, ResNet, Inception, and ViT—we discerned consistent patterns in their responses to mirror images, closely mirroring human perceptual tendencies.

The convergence of DNN behavior with human perception underscores the efficacy of these models in approximating complex cognitive processes. Furthermore, our comparative assessment utilizing the confusion index not only corroborates the alignment between DNNs and human perception but also provides insights into the nuanced similarities and differences across different model architectures.

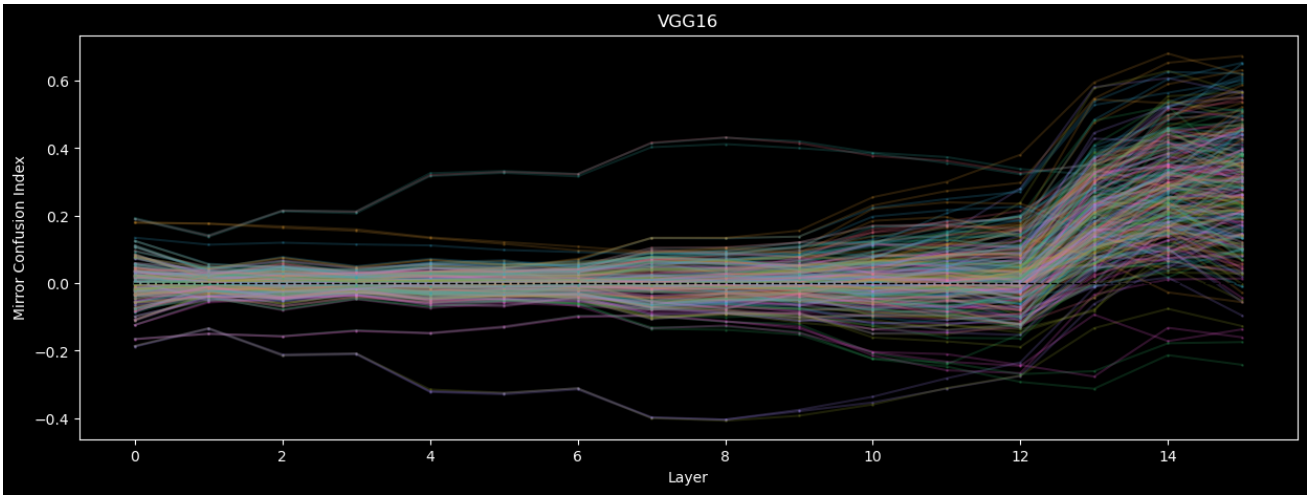

Fig. 8. Layer-wise confusion index for every image

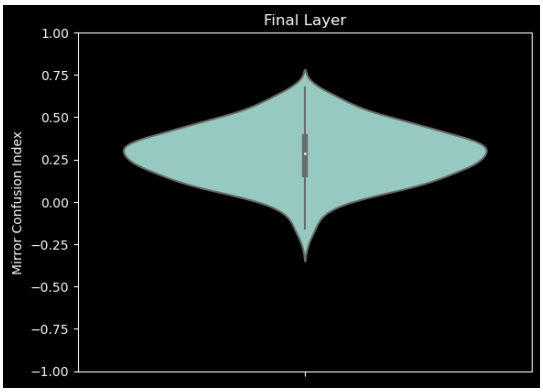

Fig. 9. VGG16 confusion indexes distribution for every stimuli

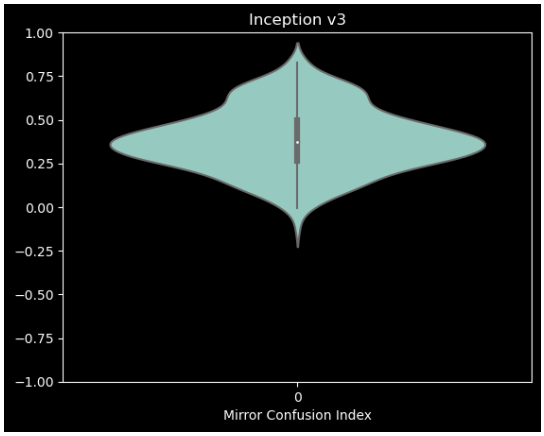

Fig. 11. Inception confusion indexes distribution for every stimuli

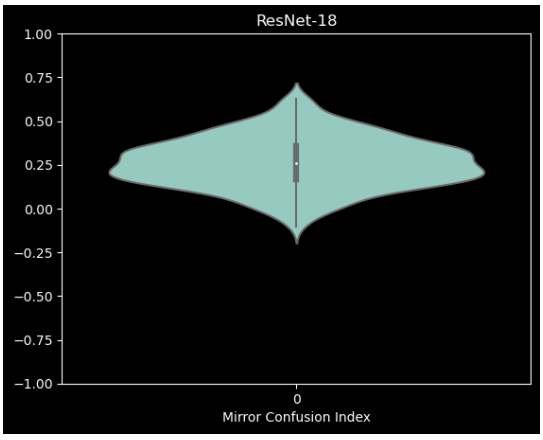

Fig. 10. ResNet confusion indexes distribution for every stimuli

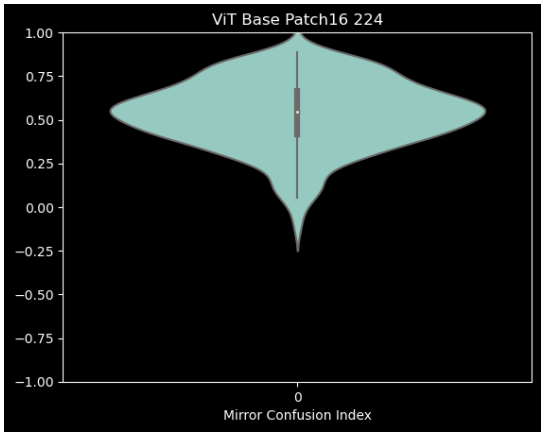

Fig. 12. ViT confusion indexes distribution for every stimuli

III. Discussion

Qualitative Analysis of Object Representations in Brains and Deep Neural Networks (DNNs)

Our study investigates the qualitative similarities and differences in object representations between human brains and DNNs trained for object recognition. We examine how certain properties manifest in randomly initialized networks, those trained for object classification, and those absent even after training. The results are summarized in Table 1 and extend across various instances of the VGG-16 network, other feed-forward neural networks, and different distance metrics.

Implications and Insights:

1) **Clarification of Model Accuracy:** Our findings shed light on when DNNs can accurately model biological vision, highlighting instances where they align or diverge from human perceptual processes.
2) **Origins of Emergent Properties:** We discern that some properties arise from the inherent architecture of DNNs, while others emerge only after training for object classification. Additionally, certain properties remain absent even post-training.
3) **Potential for Model Improvement:** Understanding these findings presents opportunities for enhancing DNN performance. By incorporating missing properties as training or architectural constraints, we can potentially enhance model robustness and accuracy.

Addressing General Concerns:

1) **Validity of Testing with Artificial Images:** Contrary to concerns, our results are consistent with traditional approaches in psychology and neuroscience that use artificial images to elucidate visual processing, highlighting the relevance of our findings.
2) **Potential Training Adaptability:** While it might be argued that DNNs could be trained to exhibit all tested properties, our findings suggest otherwise, hinting at potential limitations in learning certain relational properties by computer vision algorithms.

Key Findings:

1) **Properties in Randomly Initialized Networks:** Several properties, including correlated sparseness and relative size encoding, are present in randomly initialized DNNs, underscoring the role of network architecture in generating meaningful features.
2) **Properties in Trained DNNs:** Trained DNNs exhibit properties like Weber's law and scene incongruence, suggesting sensitivity to image regularities and contextual influences arising from object classification training.
3) **Absent Properties:** Notably, properties like 3D processing and surface invariance are absent in both randomly initialized and trained DNNs, indicating the potential emergence of these properties with additional task demands.

Future Directions:

1) **Improving DNN Performance:** Insights from this study can inform strategies for improving DNN performance by explicitly training models to produce missing properties alongside object classification tasks or exploring alternative training tasks such as navigation or agent-object interaction.
2) **Addressing Susceptibility to Adversarial Attacks:** Training DNNs to exhibit perceptual and neural properties described in our study may enhance model robustness to adversarial attacks, mitigating vulnerabilities observed in current state-of-the-art models.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.