論 文 内 容 の 要 旨

Deep Neural Networks (DNNs), an approach inspired by human brain internal mechanism, have revolutionized fields like image recognition, natural language processing, and speech recognition through their advanced feature extraction and pattern recognition capabilities. This success mainly depended on that DNNs achieve high-level feature extraction and complex pattern recognition through a multi-layered neuronal structure, enabling the extraction of deep features from data. However, despite their outstanding performance in handling these tasks, DNNs still exhibit significant limitations in robustness and interpretability. They are highly sensitive to minor changes in input data and are easily make errors when facing adversarial attacks or extreme situations, revealing their sensitivity to inputs and insufficient understanding of complex contexts. In other word, there is still a significant gap between DNNs and the human visual system.

These limitations have prompted researchers to seek inspiration from brain-like computing to improve neural network design and performance. The human brain exhibits high robustness and flexibility in processing visual information, making accurate judgments even in complex and ambiguous environments. In particular, the brain integrates contextual information and multi-level feature processing to address various visual challenges. This hierarchical processing structure and feedback mechanism provide important insights for the improvement of DNNs. Based on this purpose, research in artificial vision systems has increasingly focused on how to enhance the visual perception capabilities of models to make them more akin to the human visual system. For example, by enhancing the contextual understanding and temporal information processing capabilities of DNNs, introducing recurrent networks (RNNs) and self-attention mechanisms to simulate the feedback loops and attention regulation in the visual cortex. These advancements have improved DNNs' performance and robustness in complex visual tasks, such as object recognition and scene understanding. Moreover, models such as Generative Adversarial Networks (GANs) has also demonstrated powerful potential in data generation and understanding.

However, despite these improvements enhancing some brain-like features of DNNs, such as processing speed and accuracy, they are still limited in simulating real neurobiological functions. Achieving brain-like characteristics is not merely through

simulating a single process or mechanism; it requires a deep understanding and comprehensive simulation of various brain functions. In this context, research around brain-like computing has deepened, exploring various directions, including the study of optical illusions.

The integration of visual illusions provides a new perspective for DNN research, using brain-like mechanisms to reveal and understand the limitations of neural networks. Visual illusions serve as an intriguing tool to explore the parallels and differences between human visual perception and machine vision. These illusions often exploit the ways in which humans process visual information, revealing the underlying mechanisms of our perception. Historically, visual illusions have been used to probe the workings of the human brain, offering insights into depth perception, color constancy, and the geometrical interpretation of space. Therefore, this research topic can guide potential improvements in model optimization and training methods., by studying how DNNs handle these illusions, researchers can uncover the extent to which neural networks simulate human-like perception and where they differ, shedding light on both the capabilities and limitations of these systems.

Thus, this study delves into the simulation of human visual perception by DNNs, using a unique and comprehensive visualization approach that integrates six classical visual illusions to probe and compare the brain-like characteristics of different DNNs architectures. Specifically, depending on the human perceptual data as benchmark, we integrated visual and analytical techniques, including representational similarity analysis and class activation maps (CAM), to provide deeper explain of internal mechanism into how DNNs process visual illusions. For instance, GradCAM shows the image areas focused on by DNNs when making decisions, revealing key features that might be considered during the processing of illusions. These methods help us understand the internal workings of DNNs when dealing with visual illusions. In addition, the DNNs models we utilized in this study are both according to Brain-Score and BH-score, which are current brain-like rankings on visual pathway mapping. This study also considers the other types of DNNs, such as spatiotemporal and predictive decoding models, to explore the universality of DNNs on visual illusion completely.

Based on the proposed comprehensive interpretive visualization method, the study's specific approach is divided into four steps which respond to four chapters: firstly, verifying and testing the pre-trained DNNs' performance on visual illusions, followed by comparisons based on training with specific visual illusion datasets.

Next, differences based on the models' architectures are examined in detail, and finally, the findings are used to explore potential brain-like characteristics through fMRI experiments.

In Chapter 3, several top-ranking DNNs on Brain-Score were selected to test the Müller-Lyer illusion. The differences among the models in terms of feature attention distribution were significant. Advanced models with excellent performance in visual tasks, such as the Transformer-based ViT and Swin-T, did not exhibit visual illusions. In contrast, classic networks with single architectures like AlexNet and ResNet101 showed the illusion of line length change. This phenomenon emphasizes the differences in brain-like characteristics of DNNs, where high performance in visual tasks does not equate to brain-like characteristics. For example, in our Chapter 3 testing on five types of illusions—focusing on color, brightness contrast, length, angle, and perception—it is discussed that regarding color sensitivity, among 12 DNN models, only two align relatively well with human perception of color depth rankings. There is no regular pattern in the ranking distribution among the models, but an increase in network depth leads to changes in color ranking, indicating a change in color sensitivity, though this change is only apparent in the last module. In terms of feature focus visualization, DNNs also show significant differences, with ResNext101 recognizing the entire color rectangle and focusing on the whole area, while other models focus on partial areas. This differs from our understanding of vision; DNNs cannot comprehend color and its resulting shapes, affecting attention differences in color depth rankings. Moreover, ResNext101 does not exhibit a performance closer to the depth ranking of human subjects. Although some advanced DNNs perform well in visual tasks, they may lack the ability to handle certain human visual illusions, whereas some simpler traditional networks may more closely resemble human visual system characteristics in certain aspects.

In Chapter 4, further training on multiple models with specific datasets showed significant differences among the models. Notably, VGG19 almost did not exhibit any visual illusions during this training. The training with specific visual illusion datasets mainly aimed to develop DNNs' understanding of single physical attributes, followed by related physical attribute visual illusion tests on these trained models, such as the tilt illusion. The results showed that the performance of DNNs in visual illusions is indeed influenced by the training datasets. Among them, ResNet101 performed the best in the tests, achieving a classification accuracy of 90.28%, and excelling in recall and F1 scores. Although VGG19's feature attention distribution was similar to ResNet101, it did not exhibit any visual illusions in the tests, with an

accuracy of only 61.81%. Additionally, ResNet101's representational dissimilarity matrices (RDMs) indicated the highest representation similarity in its early modules, suggesting the importance of visual illusion responses in early visual regions (such as the V1 area). The analysis also revealed that most models performed exceptionally well on colors like "green," "spring green," "cyan," and "yellow," with many achieving 100% accuracy, but performed poorly on "blue," "magenta," and "purple." EfficientNet-B1 and ResNet101 showed higher accuracy across most colors, reflecting their potential advantage in handling natural tones, while EfficientNet-B6 and VGG19 showed lower accuracy on "orange" and "purple." In terms of strength recognition, most models were more accurate at medium strength but had challenges at extreme strengths. ResNet152 and DenseNet201 performed well across most strength levels, while ResNet101 also showed balanced capabilities at medium strength. VGG19 and PNASLarge performed poorly at extreme strength levels, and EfficientNet-B6 had limitations at low strength. Furthermore, visualization techniques like Grad-CAM revealed distinct feature trends between illusion and non-illusion stimuli across DNNs, emphasizing the complex interplay of neural and computational mechanisms in visual perception. These findings highlight the intricate processing layers in DNNs and their varying capabilities in handling visual illusions, with models like ResNet101 demonstrating superior performance across different strengths and colors of illusions.

Combining the temporal and static characteristics of the models, Chapter 5 explored the visual illusion performance of four video classification models and one predictive coding model. A new training strategy, teacher-student self-supervised learning, was proposed to fully simulate human-like learning methods to enhance the brain-like characteristics of DNNs. The results showed that the models exhibited visual illusion responses in terms of representational similarity, particularly similar to the distribution shown by previous static models. However, in GradCAM analysis, static models like AlexNet, VGG19, and ResNet101 focused more on the arrows themselves in the feature attention heatmaps, similar to the human visual system, which is strongly influenced by the direction of the arrows when perceiving visual illusions. In contrast, the video models only focused on the combination of arrows and lines. This significant difference in attention indicates that although video models have advantages in global and spatiotemporal analysis, they may be less precise than static models in capturing key visual cues directly related to visual illusions. Additionally, the study revealed that under training with Type A and Type B datasets, the four video models exhibited distinct behaviors. For instance, MViT-V1-

B consistently showed the greatest dissimilarity, indicating a significant difference in the perceived lengths of Müller-Lyer lines between the perception and control groups. S3D and R3D-18 exhibited a trend of decreasing dissimilarity with increasing labeled line lengths, suggesting varying sensitivity to line length based on the training dataset. Moreover, RDM analysis indicated that R3D-18, MViT-V1-B, and S3D displayed high similarity on the diagonal in both Type A and Type B datasets, implying that these models perceive line lengths similarly regardless of arrow orientation, akin to human visual illusions. However, Swin3D-T's irregular similarity distribution suggests it does not effectively understand the Müller-Lyer illusion. These findings underscore the models' varied capabilities in recognizing and interpreting visual illusions, highlighting the need for further refinement to enhance their brain-like characteristics.

In Chapter 6, fMRI-based experiments further explored the correlation between visual regions and visual illusions, based on the visual illusion data from Chapter 4. The results showed that the response regions were closer to the early regions of the ventral pathway, such as V1/V2, similar to the RDM distributions at different network depths in Chapter 4. This result suggests a potential relationship between DNNs and the ventral pathway in the human visual system, highlighting the importance of shallow modules in brain-like modeling of DNNs. The systematic analysis revealed significant activation differences among the ROIs under conditions of illusion and non-illusion. Notably, V2 showed a pronounced response under illusion conditions, underlining its crucial role in reaction of visual illusions, whereas V1 demonstrated stronger activation under non-illusion conditions, indicating its dominance in processing basic visual elements. As for individual perception, where significant variability in responses to illusions among participants underscored the complexity of perceptual processing in the human visual system. This variability mirrors the responses of DNNs, suggesting some commonalities in visual processing strategies between the human brain and DNNs, particularly in primary visual processing areas.

In summary, this comparative study of DNNs architectures and classical visual illusions provides important insights into the differences between human and DNNs perception. The main contribution of this study is as following:

1. This study contributes to the understanding of DNNs behavior in visual illusions and establishes methods for further examining their brain-like processing capabilities. We provide evidence demonstrating the potential brain-like advantages and limitations of DNNs.

2. By integrating neuroscientific findings into DNNs development, this work supports targeted improvements in network architecture to more closely align with human cognitive processes. Through detailed analysis and experimental insights, this research provides the reference on improving DNNs' performance in tasks requiring complex visual processing and interpretation.

3. This study reveals the strengths and weaknesses of DNNs in handling visual illusions, offering new perspectives on their potential and limitations in practical applications. For example, in fields such as autonomous driving, medical image analysis, and human-computer interaction, understanding and improving the visual perception capabilities of DNNs can significantly enhance their performance and reliability.

Finally, to enhance the brain-like characteristics of DNNs, future work needs to further set specific visual illusion datasets and design models with specific architectures, particularly focusing on the feature information of shallow modules for brain-like modeling. Through such optimizations, we can better simulate the human visual system, thereby promoting the development and application of artificial intelligence technology.