Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230212

Research on image classification leveraging deep convolutional neural networks and visual cognition

Chen Liu

College of Software Engineering, Sichuan University, Chengdu, 610065, China

2020141060049@stu.scu.edu.cn

Abstract. The field of image classification has experienced remarkable improvements with the advent of deep learning techniques, especially Deep Convolutional Neural Networks. The present study provides an extensive exploration of the junction where image classification based on Deep Convolutional Neural Networks meets human visual cognition. Utilizing the inherent ability of these networks to automatically learn hierarchical features from raw pixel data, this research examines their potential in classifying images from diverse complex datasets, emphasizing predominantly on the extensively utilized ImageNet dataset. The initial aspect of this study involves training and evaluating models based on Deep Convolutional Neural Networks on the ImageNet dataset, which comprises millions of labeled images spanning across thousands of categories. Well-established network architectures such as AlexNet, VGGNet, GoogLeNet, and ResNet are employed, and their performance in the challenging task of image classification is assessed. Rigorous experiments highlight the strengths and weaknesses of each model while emphasizing the prospects of transfer learning and fine-tuning. Following this, the interpretability of Deep Convolutional Neural Networks is explored by using visualization techniques to comprehend the learned feature representations. By visualizing activation maps and class-specific saliency maps, valuable insights are gained into the regions of interest that guide the decision-making of these models. Moreover, the correlation between the features extracted by these models and human visual attention mechanisms is examined to shed light on the focus of attention of the models. The study also addresses the difficulties that adversarial attacks, data bias, and generalization capabilities present to Deep Convolutional Neural Networks. Strategies to enhance the robustness and adaptability of the models across various domains are examined, linking these observations to human cognitive behavior.

Keywords: cognitive science, computer vision, feature extraction, semantic, image classification.

1. Introduction

The research of depth convolutional neural network network (DCNN) in the field of image classification has far-reaching theoretical and practical significance. First, the need for automatic image recognition and classification has exploded in various fields, such as autopilot, security monitoring, medical diagnosis, space exploration and so on. These applications require high accuracy and efficiency of image classification, while the traditional rule-based or statistical image processing methods can process complex, large-scale and high-noise image data, the results are often poor.

DOI: 10.54254/2755-2721/32/20230212

As a deep learning model simulating human neural network, DCNN has shown excellent performance in image recognition task. By learning a large amount of image data, DCNN can automatically extract and learn the hierarchical features of the image, and carry out effective classification. However, despite the impressive performance of DCNN, its internal working mechanism is not clear, especially the process of how to extract and integrate image features, which leads to its performance in some specific scenarios, for example, the classification of novel objects, the classification of images with complex background or subtle changes, there are still challenges. Therefore, the research of DCNN in image classification can not only promote the progress of image processing technology and solve the problems in practical application, but also can compare and imitate the visual processing mechanism of human brain, to deepen our understanding of human visual cognitive processes and promote the cross-study of cognitive science, neuroscience and artificial intelligence.

2. Relevant theories and techniques

2.1. Feature integration theory

The Feature Integration Theory is a cognitive psychology concept developed by Anne Treisman and Garry Gelade in 1980. This theoretical framework provides an understanding of how different characteristics are combined during the perception and recognition of objects.

According to this theory, the initial stages of the perceptual process involve a primary analysis of external stimuli, including color, shape, and motion. These attributes undergo separate processing, then are relayed to the brain's early visual processing stage, referred to as the preattentive stage. Attention plays a crucial role within this theory. Controlled by attention, the information corresponding to diverse features is amalgamated, thereby forming a complete object perception. This integration occurs within the later stages of the visual system, recognized as the focused attention stage.

Feature Integration Theory underscores the importance of two stages: the independent processing of characteristics and their subsequent integration. It suggests that when there are substantial differences between the features of external stimuli, they can be easily consolidated into a comprehensive object during the integration phase. Conversely, if these features exhibit similarity, the integration procedure might be disturbed, potentially leading to errors in perception or recognition difficulties.

One classic experiment related to Feature Integration Theory is the "Feature Search" experiment proposed by Anne Treisman and Garry Treisman. In this experiment, participants are asked to find a specific target object within a group of objects with the same features, for example, finding a green circle among a group of red circles. The results show that the search task is easier when there is a significant feature difference between the target and the background. As the feature similarity between the target and the background increases, the search task becomes more difficult.

Feature Integration Theory is significant for understanding the mechanisms of human perception and visual cognition. It reveals the roles of feature analysis and integration in visual processing, as well as the modulation of attention in the integration process. This theory has broad applications in cognitive psychology, computer vision, and human-computer interaction.

2.2. Principles of human-computer interaction

The Principles of Human-Computer Interaction (HCI) provide guidelines and concepts for designing user-friendly and effective interfaces between humans and computer systems. These principles aim to enhance the usability, efficiency, and overall user experience of interactive systems. Here are some key principles:

User-Centered Design: The design process should revolve around the needs, goals, and abilities of the users. It involves understanding the users' tasks, behaviors, and preferences, and incorporating their feedback throughout the design and development stages.

Visibility and Feedback: The system should provide clear and immediate feedback to users, informing them about the system's state and the outcome of their actions. Visual cues, progress

DOI: 10.54254/2755-2721/32/20230212

indicators, and error messages are examples of providing feedback to enhance user understanding and control.

Consistency: The interface elements and interactions should be consistent throughout the system to minimize cognitive load and facilitate learning. Consistency includes using standardized design patterns, terminology, and navigation structures to create a predictable and familiar user experience.

Simplicity: The interface should be kept simple and intuitive, avoiding unnecessary complexity. This principle emphasizes removing unnecessary elements, reducing cognitive overload, and presenting information and functionality in a clear and understandable manner.

Flexibility and Efficiency: The system should provide flexibility for different user preferences and support various workflows. It should allow users to customize settings, provide shortcuts, and streamline repetitive tasks to enhance efficiency and productivity.

Error Prevention and Recovery: The interface should be designed to prevent errors through clear instructions, proper affordances, and validation mechanisms. Additionally, it should provide users with the ability to undo or recover from errors and provide meaningful error messages to guide users in resolving issues.

Accessibility: The interface should be accessible to users with different abilities, including those with disabilities. Design considerations such as proper contrast, keyboard accessibility, alternative text for images, and assistive technology support are essential for inclusive design.

Learnability: The system should be easy to learn and navigate, enabling users to quickly understand its functionality and features. Clear instructions, onboarding processes, and intuitive interactions contribute to the learnability of the system.

These principles, along with user research and iterative design processes, guide HCI professionals in creating interfaces that meet users' needs and enhance their overall interaction with computer systems.

2.3. Deep convolutional neural network

Deep Convolutional Neural Networks (DCNNs) have revolutionized the field of computer vision, particularly in image classification tasks. Several popular DCNN architectures have been developed over the years, including AlexNet, VGGNet, GoogLeNet, and ResNet. Here's a brief overview of these architectures: As shown in Figure 1.

Figure 1. Deep neural network and its development [1].

AlexNet: AlexNet, proposed by Alex Krizhevsky et al. in 2012, played a pivotal role in popularizing deep learning for image classification. It was the winner of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2012. AlexNet consists of eight layers, including five convolutional layers and three fully connected layers. It introduced the concept of using Rectified Linear Units (ReLU) as activation functions, local response normalization, and dropout regularization. As shown in Figure 2.

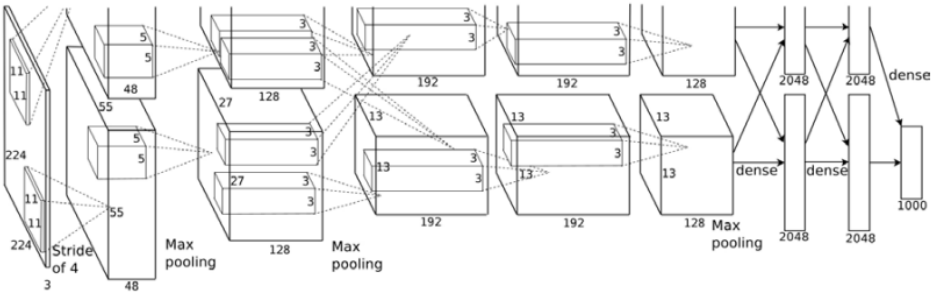

Figure 2. Alex Network structure diagram [2].

DOI: 10.54254/2755-2721/32/20230212

VGGNet: VGGNet, developed by the Visual Geometry Group at the University of Oxford in 2014, is known for its simplicity and depth. It is characterized by its uniform architecture, where 3x3 convolutional filters are stacked repeatedly, and max-pooling is performed after every two or three convolutional layers. VGGNet achieved excellent performance in the ILSVRC 2014 competition and is widely used as a baseline network for many computer vision tasks. As shown in Figure 3.

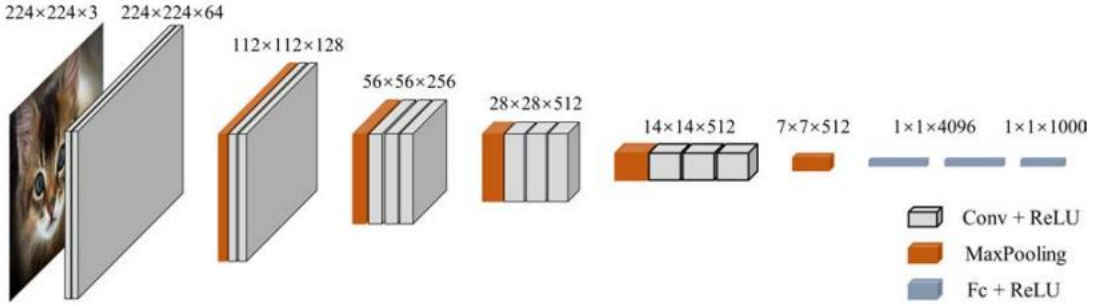

Figure 3. VGG Network structure diagram [3].

GoogLeNet (Inception): GoogLeNet, introduced by Szegedy et al. from Google Research in 2014, introduced the concept of the Inception module. It aims to address the trade-off between network depth and computational efficiency. The Inception module performs multiple convolutions with different filter sizes and concatenates their outputs, allowing the network to capture features at different scales. GoogLeNet was the winner of the ILSVRC 2014 competition and demonstrated state-of-the-art performance with a significantly reduced number of parameters. As shown in Figure 4.

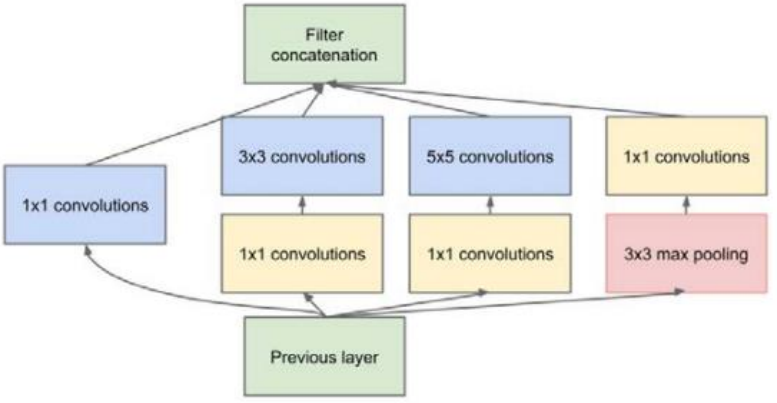

Figure 4. GoogLe network structure diagram [4].

ResNet: ResNet (Residual Network), proposed by Heetal. in 2015, introduced the concept of residual connections to address the degradation problem in deep networks. Residual connections allow information to bypass certain layers, enabling the network to learn residual mappings. This architecture significantly improved the training and performance of very deep networks. ResNet won the ILSVRC 2015 competition and has been widely adopted in various computer vision tasks. As shown in Figure 5.

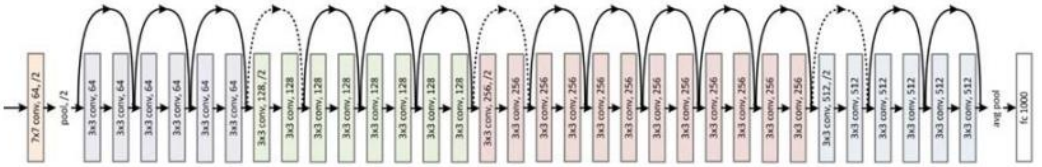

Figure 5. ResNet34 Network structure diagram [5].

DOI: 10.54254/2755-2721/32/20230212

These DCNN architectures have had a significant impact on the field of computer vision, demonstrating exceptional performance in image classification, object detection, and other related tasks. They serve as foundational models and have paved the way for subsequent advancements in deep learning and computer vision research.

3. System analysis and application research

3.1. Experimental framework

The experimental framework for image classification based on DCNN and visual cognition typically involves the following components: As shown in Figure 6.

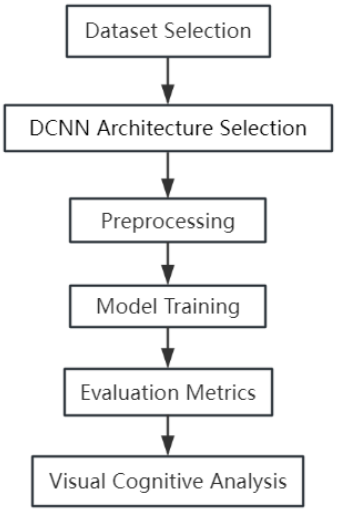

Figure 6. Experimental process diagram [6].

3.2. Data set

The ImageNet dataset is an open, publicly available dataset that allows researchers and developers to use it freely for image classification tasks.

The images in ImageNet are from real-world scenes and contain a variety of objects, backgrounds, and perspectives. Compared with the synthetic data set, ImageNet images are closer to the actual application scene, which makes the trained model have better generalization ability. As shown in Figure 7.

Figure 7. The ImageNet dataset [7].

DOI: 10.54254/2755-2721/32/20230212

3.3. Research methods/approach

DCNN Architecture Selection: AlexNet, VGGNet, GoogLeNet, or ResNet, based on the specific requirements of the experiment. Consider factors like model complexity, performance, and availability of pre-trained models. As shown in Figure 8.

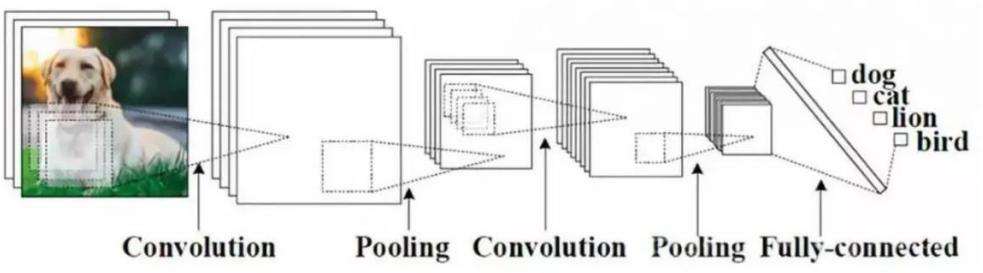

Figure 8. The convolutional neural network [8].

The formula for convolution is shown in Formula (1):

$$X_j^l = f(\sum_{i=m_j} k_{ij}^l * X_i^{l-1} + b_j^l) \quad (1)$$

The formulas for average and maximum pooling are shown in (2) and (3):

$$X_j^l = f(\frac{1}{m}\sum_{i=m_j} X_i^{l-1} + b_j^l) \quad (2)$$

$$X_j^l = f(max_{i=mj} X_i^{l-1} + b_j^l) \quad (3)$$

Preprocess the dataset and images to ensure uniformity and compatibility with the chosen DCNN architecture. Here are some common preprocessing techniques: Image Resizing: Resize the images to a consistent size to ensure uniformity in the input data. This step is necessary because images in a dataset may have different resolutions and aspect ratios. The input image size of all models was scaled to 224 × 224, and then the input feature extraction model was used, and the feature extraction model was pre-trained on ImageNet in PYTORCH model library to extract the full-connection layer, the output size of AlexNet is 256 × 6 × 6, the output size of VGG is 512 × 7 × 7, the output size of GoogLeNet is 1024 × 1 × 1, and the output size of ResNet is 512 × 1 × 1.

Data Normalization: Normalize the pixel values of the images to bring them within a specific range or distribution. Common normalization techniques involve scaling the pixel values between 0 and 1 or standardizing them using mean and standard deviation. Normalization helps to mitigate the impact of varying intensity levels and facilitates stable model training.

Data Augmentation: Augment the dataset by applying transformations to the images. Data augmentation techniques can include random rotations, translations, flips, and crops. By introducing variations in the training data, data augmentation helps improve the model's generalization capability, reduces overfitting, and increases the effective size of the dataset.

Noise Reduction: Apply noise reduction techniques, such as Gaussian blurring or median filtering, to remove or reduce image noise. This can enhance the clarity of the images and reduce the influence of noise during feature extraction and classification.

Model Training: Train the DCNN model using the labeled dataset. This typically involves feeding the preprocessed images through the network, adjusting the model's weights and parameters using optimization algorithms like stochastic gradient descent (SGD) or Adam, and iteratively updating the model to minimize a predefined loss function.

DOI: 10.54254/2755-2721/32/20230212

3.4. Experimental results and analysis

3.4.1. Evaluation metrics. Model Training: The training of the Deep Convolutional Neural Network model commences with the ImageNet dataset, comprising millions of images spanning a wide variety of categories. Each training cycle, also known as an epoch, involves processing a batch of images with the model. The model's weights get updated via an optimization algorithm known as stochastic gradient descent, and subsequently, the model's performance is assessed based on the training data. As the training process unfolds, the model's accuracy and the associated loss at the end of each epoch are recorded. This allows for tracking the model's learning progress over time. To better visualize the progression of the training, a graph is plotted. The horizontal axis represents the number of training epochs, while the vertical axis shows the corresponding training accuracy and loss values. This graphical representation helps in understanding the relationship between the number of iterations and the performance of the model. As shown in Figure 9.

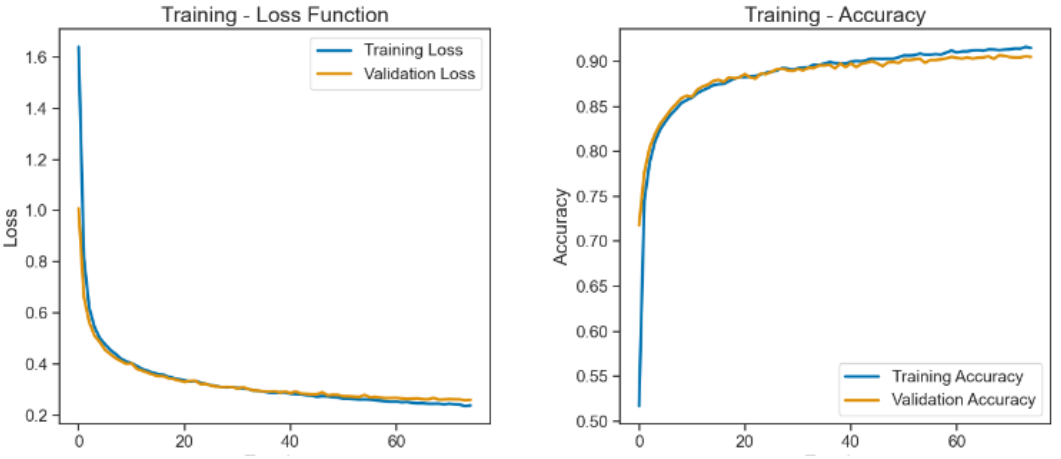

Figure 9. The relationship between training accuracy and loss [9].

3.4.2. Visual cognitive analysis. Incorporate visual cognitive analysis techniques to understand the model's behavior and relate it to human visual cognition. compare the model's predictions with human perception and decision-making.

To delve into the interpretability of DCNNs and gain insights into the learned feature representations and attention focus, several visualization techniques can be employed. Some commonly used techniques include:

Activation Maps: Activation maps, also known as feature maps or activation patterns, visualize the response of individual neurons or filters in the DCNN to specific input stimuli. By visualizing the activation maps of different layers, researchers can identify which regions of the input image trigger higher activations in the model, providing clues about the learned features and the model's perception of different visual patterns.

Class Activation Maps (CAM): Class activation maps highlight the regions of an input image that contribute most to the model's decision for a specific class. CAMs are derived from the gradients of the output with respect to the feature maps, providing insights into which image regions are relevant for the model's classification decision. As shown in Figure 10.

DOI: 10.54254/2755-2721/32/20230212

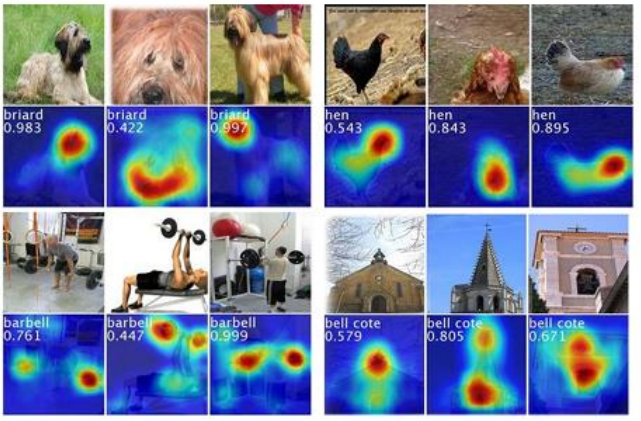

Figure 10. The CAMs of four classes from ILSVRC [10].

Saliency Maps: Saliency maps highlight the most salient or informative regions of an input image that contribute to the model's prediction. These maps are computed by measuring the sensitivity of the model's output to small changes in input pixels. Saliency maps can reveal which parts of an image receive the most attention from the model during classification. As shown in Figure 11.

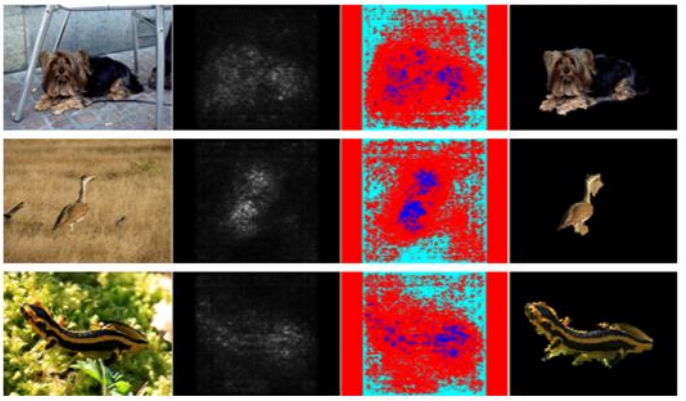

Figure 11. Object recognition and segmentation in image [11].

4. Challenges

Interpretability and explainability: DCNNs are black-box models, making it difficult to interpret and understand their internal workings and decision processes. This poses a challenge for visual cognition research in image classification. From the perspective of understanding human visual cognition and decision-making, mapping and explaining the outputs of DCNNs to human cognitive processes remain an open question.

Data bias and imbalance: DCNNs often require large amounts of labeled data for training in image classification. However, real-world datasets often exhibit class distribution imbalance and labeling errors. This can lead to DCNNs learning biases toward common classes or poor recognition performance for rare classes. Addressing data bias and imbalance is a challenge in image classification research.

Adversarial attacks and robustness: DCNN models in image classification can be susceptible to adversarial attacks, where slight perturbations to input images cause the model to produce incorrect classification results. This suggests differences between the processing of visual information by DCNNs and human visual cognition. Improving the robustness of DCNNs and their ability to resist adversarial attacks to better simulate human visual cognition is an important research direction.

Transfer learning and generalization: DCNN models may achieve good performance on the training set but still face challenges in generalizing to unseen data. How to transfer the learning capabilities of

DOI: 10.54254/2755-2721/32/20230212

DCNNs from one task to another and how to maintain performance stability in different domains or environments are issues that need to be addressed.

These challenges require further research and exploration, combining DCNNs with visual cognition research to improve the performance of image classification and understand the mechanisms of human visual cognition. Addressing these challenges will drive advancements in computer vision and cognitive science, facilitating the development of more effective and interpretable methods for image classification.

5. Conclusion

This study presents an exhaustive exploration of image classification leveraging Deep Convolutional Neural Networks (DCNNs), and how it aligns with human visual cognition. DCNNs are renowned for their ability to learn hierarchical features autonomously from raw pixel data. The study evaluates their performance using the challenging ImageNet dataset, employing well-known DCNN architectures such as AlexNet, VGGNet, GoogLeNet, and ResNet. The results underscore the potential of these networks in complex image classification tasks and transfer learning scenarios. Interpretability forms a crucial part of this study. Various visualization techniques are utilized to interpret the learned feature representations. Through visualizing activation maps and class-specific saliency maps, invaluable insights into the regions of interest that influence the model decisions are gathered. These visualizations illuminate the features learned by the model, revealing the model's perception of visual patterns.

The study additionally tackles challenges associated with adversarial attacks, data bias, and generalization capabilities in DCNNs. Methods to enhance model robustness and adaptability across multiple domains are investigated, striving to align the models' behavior with the intricacies of human cognitive processing.

In summary, the research provides deeper insight into image classification based on DCNNs and its relationship with human visual cognition. It exemplifies the capabilities of DCNNs in large-scale image classification tasks and their potential in transfer learning scenarios. The use of visualization techniques leads to valuable insights into the learned features and the model's attention focus. The visual cognitive analysis offers evidence of the models' strengths and limitations compared to human perception.

The conclusions drawn from this study serve as a stepping stone for future advancements in the field of computer vision research. By bridging the gap between machine-based image understanding and the complexities of human visual cognition, this research sets the groundwork for the development of more interpretable and human-aligned image classification models. As the landscape of deep learning continues to evolve, this research contributes significantly to the quest for more robust, interpretable, and human-like Artificial Intelligence systems in the field of image classification.

References

[1] Yan Jianpu, Y. (2022). Research on brain-computer hybrid intelligent computing method for image classification task [Dissertation]. Xidian University.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS).

[3] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of International Conference on Learning Representations (ICLR), Boston.

[4] Szegedy, C., Liu, W., Jia, Y. Q., & others. (2015). Going deeper with convolutions. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston.

[5] He, K. M., Zhang, X. Y., Ren, S. Q., & others. (2016). Deep residual learning for image recognition. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas.

DOI: 10.54254/2755-2721/32/20230212

[6] Guangzhu, X., Zequn, Z., Silu, Y., & others. (2021). Flower Image Classification system based on lightweight deep convolutional neural network. Data Acquisition and Processing, 36(4), 756-768. https://doi.org/10.16337/J. 1004-9037.2021.04.014.

[7] Huiyong, W., Chunjie, X., Xiaoming, Z., & others. (2020). Image correlation measure based on DCNN classification. Computer Applications Research, 37(2), 625-629. https://doi.org/10.19734/J. ISSN. 1001-3695.2018.04.0487.

[8] Chen, Y., Argentinis, J., & Weber, G. (2016). IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. Clinical Therapeutics, 38(4), 688−701.

[9] Zhou, A., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). arXiv:1512.04150.

[10] Russakovsky, J., Deng, H., Su, J., Krause, S., Satheesh, S., Ma, Z., Huang, A., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & FeiFei, L. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision.

[11] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034.