Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

A COMPREHENSIVE REVIEW OF CNN ARCHITECTURES FOR IMAGE RECOGNITION: ADVANCES AND OPEN CHALLENGES.

Sohail Ahmed Memon*[1], Israr Ahmed[2], Mashooque Ali Mahar[3], Ghulam Ali Alias Atif Ali Memon[4], Shereen Fatima[5]

*[1,2]*Department of Mathematics, Shah Abdul Latif University, Khairpur*
[3]*Institute of Computer Science, Shah Abdul Latif University, Khairpur*
[4]*COPELABS, Universidade Lusofona, Lisbon*
[5]*Department of Computer Science, Shaikh Ayaz University, Shikarpur*

*[1]suhail.memon@salu.edu.pk, [2]israr.memon@salu.edu.pk, [3]mashooq.mahar@salu.edu.pk, [4]ghullamali@gmail.com, [5]shereen.bhatti@saus.edu.pk

DOI: https://doi.org/10.5281/zenodo.15341574

Keywords
Convolutional Neural Networks (CNNs); CNN Architectures, Image Recognition, Vision Transformers; Challenges in CNNs

Article History
Received on 25 March 2025
Accepted on 25 April 2025
Published on 05 May 2025

Corresponding Author: *
Sohail Ahmed Memon

Abstract

Convolutional Neural Networks (CNNs), considered revolutionary, have transformed the field of computer vision, aiding exceptional advancements in image recognition tasks. With their ability to automatically learn spatial hierarchies of features, CNNs have become the backbone of most innovative image recognition systems. From their inception with LeNet in the late 1990s to the latest revolutions like Vision Transformers (ViTs), CNN architectures have undergone significant evolution. This paper provides a comprehensive review of the key developments in CNN architectures, focusing on their impact on image recognition performance, the challenges they face, and the potential future directions.

INTRODUCTION

In recent years, the Convolutional Neural Networks (CNNs) have emerged as an advanced and vital technology in computer vision for image classification tasks [1]. The CNNs were brought into computer technology in late 1980s and early 1990s, The LeNet was the first CNN architecture which was introduced by Yann LeCun. It was initially designed for handwritten digits recognition [1].

The CNNs became very well-known when AlexNet got success in the challenge called ImageNet Large Scale Visual Recognition (ILSVRC). After that, the CNNs led to a boom in deep learning research [2]. This was a new revolution in the field of computer vision, where CNNs proved to be the main pillar of most advanced image recognition systems.

The CNNs are devised to learn spontaneously hierarchical representations of features from input images. CNNs work differently where other methods employ manual engineering for features extraction, CNNs use convolutional filters for the extraction of features at different stages of abstraction, this range fall between low-level and high-level patterns like edges from the representation of objects and scenes.

The attribute of hierarchical feature learning ability makes CNNs more robust and powerful for the implementation in complex visual tasks which

comprise image classification, semantic segmentation and object detection [3].

The architecture of CNNs has been through extensive evolution. Initial CNNs, for example LeNet, were closely modest models which were based on small-scale image detections. However, the beginning of AlexNet achieved an important breakthrough, implementing innovative techniques, such as dropout regularization, Rectified Linear Unit (ReLU) activation, and GPU speed. The revolutionary performance of AlexNet in the ImageNet competition opened the doors of the development of strong and more complex CNN architectures, which in turn boosted the field of deep learning more advanced [4].

The introduction of AlexNet further enhanced the curiosity and the CNN architectures improved in terms of performance and efficiency. A CNN architecture called VGGNet [5] which comprised multiple convolutional layers and network depth with small 3×3 filters proved more impactful. Although, this architecture came up with computational costs but achieved notable results on ImageNet. Compared to VGGNet, the GoogleNet (Inception) [6] appeared with a more efficient technique by implementing multi-scale convolutions within a single layer. This architecture did not require the computational complexity in extracting the features at various levels of abstraction.

The CNN architecture called Residual Networks (ResNets) [7] came up with the idea of residual learning by implementing the identity mappings which excluded the concept of layers. The ResNets lessened the problem of vanishing gradients that mostly previals in deep networks. This revolution proved helpful in training of very deep architectures, including hundreds and thousands of layers, leading to very efficient performance in image recognition tasks. ResNets presented the potential depth as an important factor in increasing the performance of CNNs when properly accomplished [8].

Over the past few years, a significant increase and advancements have been observed in CNN architectures, with a special attention to address the challenges like model size and computational efficiency. The architecture of such an attention is the DenseNet [9], launched with increased gradient flow, with dense connectivity patterns and parameter efficiency. MobileNet [10] and EfficientNet [11] were introduced for mobile and embedded devices with a special focus on model optimization by employing the technique of depth-wise separable convolutions and neural architecture search, respectively. These architectures have presented outstanding performance. These developments reduced computational complexity and enhanced the implementation of CNNs to environments with limited resources.

Notwithstanding, many important advancements in CNN architecture, many challenges still exist. The innate black-box nature of CNNs with obscure decision-making stages obstructs the trust and interpretability. Besides, the need of large annotated datasets hinders the usage in domains with limited data. In addition to this, CNNs are observed to be vulnerable to adversarial attacks, where input images are manipulated with a little change. Such limitations can lead to erroneous predictions and obstruct the highly-needed research to increase efficiency, transparency and robustness that CNN models become more adaptable.

This review broadly explores the evolution of CNN architectures, explaining their applications in image recognition. The discussions are given on the key advancements and stubborn challenges in evolution and applicability of CNNs. This work aims to explain the trajectory of CNN research and the implications in various areas.

1. CNN Architectures

There are similarities between CNNs and traditional Artificial Neural Networks (ANNs). Both are comprised of interconnected neurons which learn and adapt through training. This section explores the evolution of CNN architectures. Convolutional Neural Networks (CNNs) share fundamental similarities with traditional Artificial Neural Networks (ANNs). Both consist of interconnected neurons that learn and adapt over training. Neurons work similar in both networks, prepare and process input data, perform operations such as multiplying inputs by weights and applying activation functions. While CNNs are excellent in image recognition.

The architecture of CNNs is specifically designed to handle images where features are pre-processed through convolutional, pooling, and fully connected

layers. These layers are basically stacked and stacking originates the CNN model, then it becomes able to learn complex patterns within images [12]. A complete CNN architecture can be seen in Figure 1.

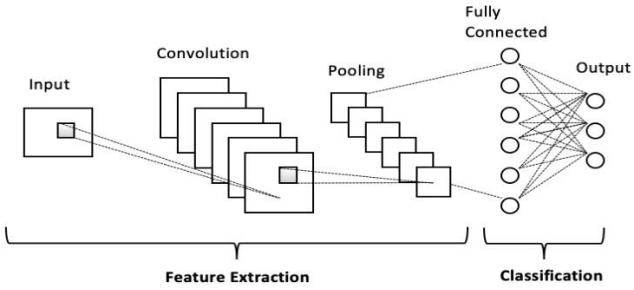

Figure 1: A fundamental Convolutional Neural Network (CNN) architecture, presenting the sequential flow from input image from feature extraction layers (convolution and pooling) to classification through fully connected layers [13].

The CNN architecture shown in Figure 1 is has five stages. At the first stage, there is an Input layer which gets raw data by feeding an image that is stored as pixel values. Convolution is the second layer which extracts features from the input image. This layer uses filters which are small matrices and detects patterns from the image. The patterns are comprised of edges, corners, and textures. Pooling is the third layer, it subsamples the output of the convolution layer, preserves the useful features and reduces dimensionality. Pooling layer is helpful in minimizing computational cost and making the network more efficient. Fourth layer is fully connected layer, works as a traditional neural network by connecting all neurons to each other. Extracted features in previous layers are forwarded to this layer then it combines them all to push to the final output for the classification or object detection. The output layer is the final layer of the network, which produces the prediction or classification result.

Early CNN Architectures

The CNNs had first emerged in the late 1980s and early 1990s. The starting was made by Yann LeCun and colleagues [14] by introducing LeNet. This revolutionary architecture used the approach of applying neural networks to visual data. The CNNs in their starting were relatively simple, such as LeNet but laid down the core techniques for recent sophisticated models. Early architectures of CNNs proved successful in tasks such as handwritten digit recognition which proved the potential of neural networks to handle complex image processing challenges.

Another architecture LeNet–5 was introuduced by Yann LeCun in 1998 [1] which was initially designed for handwritten digit recognition on the MNIST dataset. The design of LeNet–5 consisted of sequential layers of convolution, pooling, and fully connected operations. The worth of convolutional layers in extracting relevant image features and pooling layers for reduction in dimensionality was demonstrated by this pioneering model, making the CNN more robust for advancements.

LeNet had its initial application in optical character recognition, particularly identifying handwritten digits for the postal service's zip code automation. One of its remarkable achievement in classifying the MNIST dataset proved the power of CNNs for dealing complex pattern recognition challenges. The applications of LeNet increased for a variety of datasets other than digit recognition. Also, the LeNet significantly enhanced the efficiency within the financial sector, Its ability for recognizing handwritten characters was instrumental in automating the processing of bank checks and other documents. The architecture AlexNet [2] introduced by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton in 2012 [2] was considered the ground-breaking CNN, it revolutionized computer vision. The AlexNet proved itself in deep learning for image recognition by achieving a substantial margin of victory in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in the year 2012. The AlexNet CNN architecture was based on eight layers, first five convolutional layers followed by three fully connected layers. This architecture is presented in Figure 2. Key innovations included the adoption of ReLU activation functions, dropout regularization, and data augmentation to enhance

model performance. Moreover, AlexNet leveraged GPU acceleration to efficiently train on the massive ImageNet dataset.

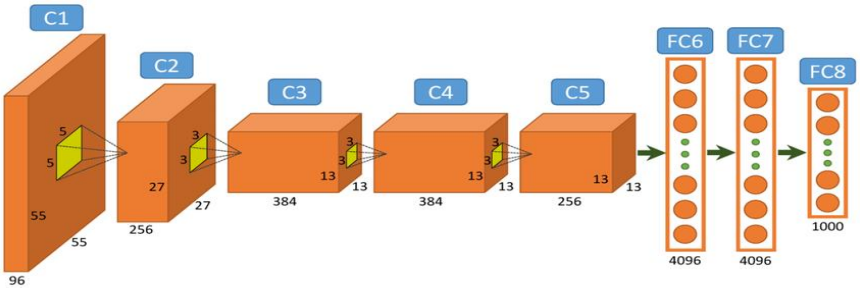

Figure 2: AlexNet convolutional neural network, highlighting the key components: convolutional layers (C1-C5), pooling layers (implied), and fully connected layers (FC6-FC8). The dimensions of feature maps and the number of neurons are shown at each layer [15].

The AlexNet architecture is comprised of mainly 8 layers, there are five convolution layers (C1-C5) and three fully connected layers (FC6-FC8). The illustration of eight layers is given in Figure 2. The AlexNet has been incorporated in various computer vision by dominating the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). With the objective of classifying images into one of 1,000 categories, AlexNet with the power of deep learning laid its best performance for object classification. This successful performance featured the potential of CNNs to outshine over traditional computer vision techniques, paving the way for advancements in a variety of image-related applications. The implementations of AlexNet enhanced and it was used to medical image recognition tasks. The applications of AlextNet have been seen in various medical image related tasks, for example, radiology scans and tumours. After the notable success of AlexNet, the scientists and researchers considered to produce CNN architectures with more depth and they developed the VGGNet. The researchers from Visual Geometry Group introduced the VGGNet at the university of Oxford. The homogeneous architecture of VGGNet is based on deep convolutional networks with 16 to 19 layers.

The VGGNet balances between increasing network depth for complex feature extraction and keeping manageable parameters, for which it uses 3×3 convolutional filters throughout. Primarily, the research on the VGGNet wat to highlight the relationship between network depth and accuracy performance in large-scale image classification tasks. This design proved highly effective, leading to a runner-up finish in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

The structure diagram, given in Figure 3, presents VGGNet16 architecture. The network consists of 16 layers, including 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. The dimensions of the feature maps are shown at each layer, along with the type of operation performed (convolution, pooling, or fully connected). The final layer is a softmax layer that produces probabilities for each class.

Besides the VGGNet16, VGGN19 is also a type of VGGNet (CNN) which has more depth of the network including cascading of convolution layers. Table 1 best describes the VGGNet16 presented in Figure 3.

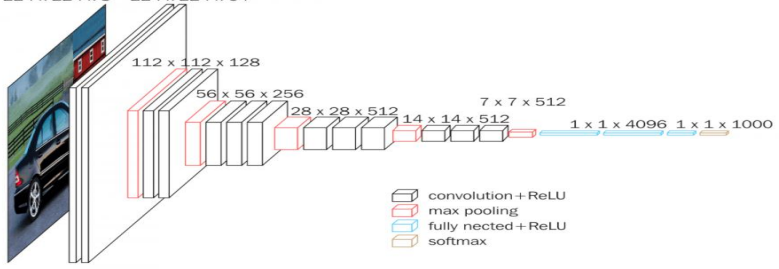

Figure 3: The architecture of a VGG-16 convolutional neural network (CNN)[16], [17].

In Figure 3, the network takes a 224x224x3 image as input, meaning the image has a width and height of

224 pixels and three color channels (RGB). Each layer, in five convolutional layers, applies 3x3 filters to the input image or the output of the previous layer, extracting features. The number of filters doubles at each convolutional layer, starting from 64 in the first layer and increasing to 512 in the last layer. After each convolutional layer, there is a pooling layer. The pooling layers subsample the feature maps by a factor of 2, reducing their spatial dimensions while preserving important information. This helps to reduce computational cost and make the network more invariant to small variations in the input image. The final three layers are fully connected layers. These layers combine the features extracted by the convolutional and pooling layers into a single vector. The first two fully connected layers (FC6 and FC7) have 4096 neurons each, while the final layer (FC8) has 1000 neurons, corresponding to the number of class labels. A softmax activation function in the output layer produces probabilities for each class and the final output with the highest probability is considered final prediction.

The VGGNet has wider applications in computer vision focusing on image classification tasks. Its efficiency to obtain complete hierarchical features has been effective in fine-grained classification, the examples of which are species identification within natural images. Due to the its resourcefulness, the VGGNet has become a cornerstone in computer vision [5]. Moreover, the VGGNet was adopted in field of neural style transfer and for its ability of extracting complicated content and style representations from images, it proved powerful in the artistic endeavour. By using the features of VGGNet, the researchers and artists can seamlessly integrate and blend the content of one image with style of another, appealing and visually striking artistic compositions [18].

The GoogleNet, which is also called Inception, is an advancement of traditional CNN architectures. The GoogleNet is comprised of innovative Inception modules and it was introduced by Szegedy et al. [6] in 2014. The GoogleNet employed parallel convolutional filters of varying sizes (1x1, 3x3, and 5x5). Additionally, the pooling operations were employed to capture features at multiple scales. Such multi-scale technique, combined with efficient parameter usage, performed better than earlier CNNs like VGGNet and AlexNet. GoogleNet marked its victory in the 2014 ILSVRC competition where it became a leading CNN architecture in image recognition.

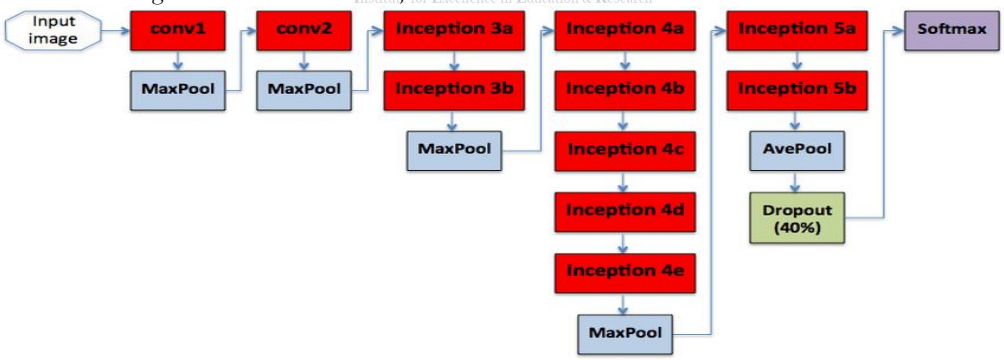

Figure 4: The architecture of a GoogleNet convolutional neural network (CNN) [19].

The architecture of GoogleNet is presented in Figure 4. An image is fed in the input layer and then using filters the features are extracted from the input image at the convolutional (Conv1, Conv2) layers. After that inception modules work for the extracted data, these modules are designed to capture features at multiple scales. The inception modules work alongside pooling operations. This is a parallel working of different-sized (Inception 3a, 3b, 4a, 4b, 4c, 4d, 4e, 5a, 5b) inception modules and pooling operations. The network learns complex features and improves accuracy by such operations. Pooling layers (MaxPool) obtain key information by reducing their spatial dimensions and forming feature maps. This process helps in reducing computational cost and making the network more adaptable to small variations in the input image. Average Pooling (AvePool) layer obtains the mean of the feature maps and produce fix-sized output. The dropout layer is used as a regularization technique to prevent

overfitting. Overfitting happens when the model performs well on training data but poorly performs for the unseen data, for example validation or test data. The dropout layer sets to zero a fraction of neurons during the training to stop overfitting of the model. A fully connected layer combines all the process from previous layers and is considered as the final layer. This layer uses the softmax function to form output into probabilities for each class.

The GoogleNet is much more capable such as its ability to learn features at multiple scales proved precious for object detection. Its accuracy of object detection enhanced for the complex scenes when this architecture was integrated into frameworks like Faster R-CNN [20]. Also, The GoogleNet outperformed for the tasks such facial recognition in which it captured all the tones and traces of facial expressions, lighting conditions and poses. Such speciality of GoogleNet has made this architecture a cornerstone in facial recognition applications.

1.1 Recent Advancements in CNNs

The CNNs have been the basis of computer vision since long, advancing in image recognition, segmentation and object detection. The innovative architectures like Vision Transformers (ViTs) are truly the recent advancements in image recognition [21]. The ViTs process images as sequences of patches focusing the self-attention mechanism and such ability of transformers from natural language processing was adapted by ViTs. This technique is unlike traditional CNNs which only rely on convolutional layers, in contrast it effectively captures the global relationships and dependencies between image patches. Thus, the ViTs have achieved a significant victory over CNNs in image recognition tasks. These advancements are renovating the landscape of deep learning for computer vision and replacing the traditional CNN dominance [22]. The Vision Transformers (ViTs), introduced by Dosovitskiy et al. in 2020 [23], have achieved hi-tech results in image classification, particularly when trained on large datasets. The ViTs have also inspired hybrid models that combine the power of two or more architectures. For example, the Convolutional Vision Transformers (CvTs) and Swim Transformers combine convolutional layers into transformer framework to increase feature extraction and reduce computational cost. Such hybrid approaches present the combined power of convolutional operations in computer vision and the attention mechanism gain power and traction.

Another remarkable development in CNNs is the Neural Architecture Search (NAS), introduced by Zoph et al. in 2017 [24], which employs a technique of iteration to yield higher probabilities to the architectures that obtain higher accuracies. Simply, the controller learns to improve its search over time. The NASNet was the first architecture developed through the NAS and outperformed for the image classification tasks. Also, several other efficient and scalable models were produced through NAS which include EfficientNet. The EfficientNet uses the technique of compounding scaling to optimize depth, resolution, model size. Recent developments, such as RegNet [25], [26], and MobileNetV3 [27] have further refined NAS techniques to design lightweight models fit for edge devices and real-time applications.

Beyond the traditional application of CNNs in image recognition, the CNNs have remarkably achieved the results in medical imaging tasks such as tumor detection[28], [29], organ segmentation [30], [31], and disease diagnosis [32]. For example, recent studies have implemented CNNs for early detection of diseases like COVID-19 from chest X-rays and CT scans [33], [34], achieving high accuracy and aiding healthcare professionals in decision-making. Likewise, in the area of self-driving [35], CNNs are employed for object detection, scene understanding, and lane detection, supporting vehicles to navigate complex environments safely.

The advancements CNNs continue to evolve quickly, enriched by the innovations such as Vision Transformers, Neural Architecture Search and further state-of-the-art applications in diverse fields. Compared to early CNN architectures, the recent innovative models have become more suitable to achieve the best and rapid results. Such new innovative techniques which include ViTs and self-attention mechanisms. On the other hand, the NAS is pushing the boundaries of automated architecture design, producing efficient and powerful models. As CNNs and related architectures continue to advance, their impact on computer vision and beyond is likely to evolve, opening new possibilities for research and application.

1.2 . Open Challenges in CNNs

Although, the remarkable achievements of CNNs in computer vision and beyond, the CNNs face various other challenges that stop their wider applicability and reliability. These challenges vary over domains including technical, ethical and practical and dealing with such challenges is highly important for the continued development of deep learning.

1.2.1 Black-Box Nature and Interpretability

One of the most persistent challenges with CNNs is their black-box nature. While CNNs achieve high accuracy in tasks like image classification and object detection, understanding how they arrive at their decisions remains difficult. This lack of interpretability is a major barrier in high-stakes applications such as healthcare, where clinicians need to trust and understand the reasoning behind a model's predictions. Recent efforts, such as explainable AI (XAI) techniques like Grad-CAM [36] and SHAP [37], aim to shed light on CNN decision-making processes. Such techniques are considered to produce false conclusions and are unable resolve the complexity of deep learning models [38].

1.2.2 Dependency on Large Annotated Datasets

The large annotated datasets for training often require an effort that can be expensive and time-consuming and to handle such datasets, the CNNs become exhausted. This tendency of CNNs limits their use in domains with short-labelled data, for example in disease diagnosis or medical imaging. To overcome such problems, the techniques like data augmentation, transfer learning and semi-supervised learning are proposed. Still these techniques often fall short of completely addressing the data management in CNNs. The researchers are working to find direction to handle these issues by employing self-supervised learning where the unlabeled data is leveraged [39], [40].

1.2.3 Computational Costs and Environmental Impact

For training and deployment of large-scale models based on CNNs require sufficient computational resources, which result in high energy consumption and environmental issues. For example, carrying a single CNN model for training can produce large amount of carbon as several cars over their lifetimes[41]. Although, the efforts are put to develop more efficient architectures which reduce these costs and environment troubles. These models include MobileNet and EfficientNet and the suitable techniques include model pruning and quantization. However, balancing efficiency with performance is still a challenge [10], [11].

1.2.4 Vulnerability to Adversarial Attacks

The CNNS are easily open to adversarial attacks, where slight change to input images can produce incorrect or improper predictions. Such weaknesses create serious risks for safety in applications like autonomous driving and facial recognition. To reduce these risks, adversarial training and robust optimization techniques are employed to enhance model robustness. To create the strong CNNs that remain resistant to such attacks is still an open challenge [42], [43].

1.2.5 Integration of Domain Knowledge

Finally, integrating domain-specific knowledge into CNN design is an underexplored area. While CNNs excel at learning patterns from data, they often fail to incorporate prior knowledge from experts, which could improve generalization and reduce data requirements. Hybrid models that combine CNNs with symbolic reasoning or physics-based constraints are emerging as a potential solution, but this area is still in its infancy[44], [45].

Nevertheless, CNNs have been recognized as a backbone in computer vision and deep learning, to address open challenges will make it more sustainable and indispensable tool in various domains.

2. Methods

The methods include some considerations how to employ a specific CNN architecture to any task. The methods can be considered on the following:

2.1 Selection of Architectures

The evolution of Convolutional Neural Networks (CNNs) has been marked by the development of several groundbreaking architectures, each contributing uniquely to the field of computer vision. This review highlights key architectures selected for

their historical significance, transformative impact, and recent advancements, including LeNet, AlexNet, ResNet, DenseNet, and Vision Transformers (ViTs).

The LeNet, introduced by Yann LeCun in 1998, was one of the earliest CNN architectures, designed for handwritten digit recognition. It laid the foundation for modern CNNs by demonstrating the effectiveness of convolutional layers combined with subsampling and fully connected layers [1]. The AlexNet, proposed by Krizhevsky et al. in 2012, revolutionized the field by winning the ImageNet competition and popularizing deep learning. Its use of ReLU activations, dropout, and GPU acceleration set new standards for CNN design [2].

The ResNet architecture achieved state-of-the-art performance on ImageNet and became an inspiration for the upcoming models. It was introduced in 2015 which addressed the vanishing gradient problem skipping connections and enabling the training of very deep networks [8]. Likewise, another architecture called DenseNet which was introduced in 2017 proved efficient. This employed the technique of feature reuse by connecting each layer to every other layer in a feed-forward manner. Thus, it enhanced parameter efficiency and performance [9]. Currently, the Vision Transformers (ViTs) have outperformed compared to CNNs for image recognition. The ViTs, which adopt the transformer architecture, are considered as alternative to CNNs. ViTs use their power of self-attention mechanisms to maintain global relationships in images and thus achieve inspiring results on large-scale datasets [23]. As far as the ViTs gain recognition and challenge traditional CNN popularity, models with attention mechanisms and combined mechanisms are gaining traction. This is the ability of hybrid models that advances the CNNs from simple feature extractions to complex and highly efficient models. Thus, the models become capable of handling diverse visual tasks. Such enduring influence of ViTs is considered the real innovation in deep learning.

2.2 Evaluation Criteria

The evaluation of CNN architectures comprises various key metrics that assure their performance, efficiency and applicability. In fact, this is the process that ensures the CNNs are not only accurate but also practical for real-world deployment across diverse domains.

The primary metric accuracy is often measured in terms of benchmarks like Top-1 and Top-5 accuracy on datasets. ImageNet, ResNet and EfficeintNet architectures have obtained remarkable high standards have achieved the notable results [8], [11]. Yet, accuracy is not all that sufficient metric but computational efficiency is equally critical. Other metrics such as floating-point operations (FLOPs) and inference time are used to evaluate how well an architecture performs under resource constraints. The MobileNet and ShuffleNet, the type lightweight models, are the examples which have exceled in this domain, and are considered suitable for edge devices [10], [46].

To make models scalable is another challenge and the criteria of scalability is important. Scalability is basically how well an architecture behaves to larger datasets or complex tasks. For example, the attribute of self-attention mechanisms in ViTs, the transformers have shown strong scalability to handle high-resolution images effectively [23]. Generally, the CNNs are vulnerable to small perturbations in input data and it is essential to protect them from adversarial attacks. Therefore, the techniques like robust optimization and adversarial training are used to increase resilience [43].

The applications of CNNs in specialized domains like medical imaging or autonomous driving is deeply evaluated. The CNN models are specifically designed for these domains, for example, U-Net is designed for biomedical image segmentation. This model is a standard in medical imaging which uses its ability to handle limited annotated data [47]. Likewise, for autonomous driving domains, the architectures like YOLO and Faster R-CNN are widely used in real-time object detection [20], [48].

The evaluation of CNN architectures goes through a combined approach of various factors which include balancing accuracy, efficiency, robustness, scalability and domain-specific applicability. All that criteria ensure that models are fully adaptable and practically sound.

2.3 Analysis Framework

To comprehensively evaluate Convolutional Neural Network (CNN) architectures, we employ a comparative analysis framework that assesses each model across multiple dimensions: performance metrics, complexity, and suitability for specific use cases. This framework enables a systematic comparison of architectures like LeNet, AlexNet, ResNet, DenseNet, and Vision Transformers (ViTs), highlighting their strengths and limitations.

- **Performance metrics** such as accuracy, inference speed, and memory usage are critical for comparing architectures. For instance, ResNet achieves high accuracy on ImageNet due to its deep residual layers, while MobileNet prioritizes efficiency, making it ideal for mobile applications [8], [10].

- **Complexity** is evaluated using parameters, FLOPs, and training time, with models like EfficientNet balancing these factors through compound scaling [11], [47].

- **Suitability** for specific tasks, such as medical imaging or real-time object detection, is also considered. For example, U-Net excels in biomedical segmentation due to its encoder-decoder structure [47], while YOLO is optimized for real-time detection [48].

3. Conclusion

In this work, the Convolutional Neural Networks (CNNs) have been thoroughly reviewed and maximum CNN based architectures have been explained. The CNNs have performed remarkably in the field of image recognition, with each subsequent development and their achievements, they have pushed the boundaries. The CNNs have set new benchmarks in accuracy and performance throughout from pioneering LeNet to the transformative AlexNet, and from the depth-defying ResNet to the parameter-efficient DenseNet. More recently, an advancement of CNNs, the Vision Transformers (ViTs) have introduced a paradigm shift, by leveraging self-attention mechanisms which can outperform the traditional convolutional approaches in certain tasks. Such advancements have played a vital role in bringing dynamic models and have also expanded the applicability of CNNs in various domains which include medical imaging, autonomous driving, and real-time object detection.

Despite the high accuracy and success of CNNs, there are some challenges and the most highlighted issue among those is the interpretability of CNNs. These models often function as "black boxes," making it difficult to understand their decision-making processes. Not only this but the lack of transparency is specifically problematic in high-stakes applications like healthcare, where interpretability is crucial for trust and adoption. To address such issues, the more in-depth research on techniques is essential in future developments to make these models more interpretable. The examples of such techniques are explainable AI (XAI) methods and hybrid models that combine the symbolic reasoning. As discussed earlier, the CNNs are vulnerable to adversarial attacks and this deficiency of CNNs lead to another critical challenge that is robustness. The small perturbations in input data can result in incorrect predictions and therefore the models must be resilient and robust. The enhanced techniques like adversarial training and robust optimization are essential for their deployment in safety-critical applications like autonomous vehicles and surveillance systems. Efficiency also matters and this concern particularly exists in resource-constrained environments. Where high-cost resources are required, the architectures like MobileNet and EfficientNet have lessened this issue by reducing the computational costs. There is a need to make CNNs more energy-efficient and environmentally sustainable. Techniques such as neural architecture search (NAS), model pruning, and quantization offer promising avenues for achieving this goal.

Looking ahead, the integration of new paradigms like transformers and the exploration of hybrid models that combine the strengths of CNNs and attention mechanisms will be crucial. These innovations promise to address existing limitations while opening up new possibilities for image recognition and beyond. The ongoing development of CNN architectures will continue to drive advancements in the field, with far-reaching implications across industries such as healthcare, transportation, and entertainment.

In conclusion, while CNNs have already transformed image recognition, addressing the challenges of interpretability, robustness, and efficiency will be key to unlocking their full potential. As the field evolves, the fusion of CNNs with emerging technologies will pave the way for even more ground-breaking achievements.

REFERENCES

[1]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] I. Goodfellow, *Deep Learning*. MIT Press, 2016.

[4] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[5] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014, Accessed: Aug. 20, 2024.

[6] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[7] S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," Mar. 25, 2016, *arXiv*: arXiv:1603.08029.

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

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[10] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[11] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[12] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Dec. 02, 2015, *arXiv*: arXiv:1511.08458.

[13] S. Balaji, "Binary Image classifier CNN using TensorFlow," Techiepedia.

[14] Y. Le Cun *et al.*, "Handwritten Digit Recognition: Applications of Neural Net Chips and Automatic Learning," in *Neurocomputing*, F. F. Soulié and J. Hérault, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 303–318. doi: 10.1007/978-3-642-76153-9_35.

[15] R. Mash, N. Becherer, B. Woolley, and J. Pecarina, "Toward aircraft recognition with convolutional neural networks," in *2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS)*, IEEE, 2016, pp. 225–232.

[16] "VGGNet-16 Architecture: A Complete Guide." Available: https://kaggle.com/code/blurredmachine/vggnet-16-architecture-a-complete-guide

[17] M. M. Khodier, S. M. Ahmed, and M. S. Sayed, "Complex pattern Jacquard fabrics defect detection using convolutional neural networks and multispectral imaging," *IEEE Access*, vol. 10, pp. 10653–10660, 2022.

[18] L. A. Gatys, "A Neural Algorithm of Artistic Style," *arXiv preprint ArXiv:1508.06576*, 2015.

[19] P. Pawara, E. Okafor, O. Surinta, L. Schomaker, and M. Wiering, "Comparing local descriptors and bags of visual words to deep convolutional neural networks for plant recognition," in *6th international conference on pattern recognition applications and methods (ICPRAM 2017)*, ICPRAM, 2017.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in*

neural information processing systems, 2015, pp. 91–99.

[21] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[22] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[23] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[25] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "RegNet: Multimodal sensor registration using deep neural networks," in *2017 IEEE intelligent vehicles symposium (IV)*, IEEE, 2017, pp. 1803–1810.

[26] J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, "RegNet: Self-regulated network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 9562–9567, 2022.

[27] S. Qian, C. Ning, and Y. Hu, "MobileNetV3 for image classification," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, IEEE, 2021, pp. 490–497.

[28] M. S. I. Khan *et al.*, "Accurate brain tumor detection using deep convolutional neural network," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4733–4745, 2022.

[29] S. Kumar, R. Dhir, and N. Chaurasia, "Brain tumor detection analysis using CNN: a review," in *2021 international conference on artificial intelligence and smart systems (ICAIS)*, IEEE, 2021, pp. 1061–1067.

[30] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Medical Physics*, vol. 44, no. 2, pp. 547–557, Feb. 2017, doi: 10.1002/mp.12045.

[31] A. E. Ilesanmi, T. Ilesanmi, O. P. Idowu, D. A. Torigian, and J. K. Udupa, "Organ segmentation from computed tomography images using the 3D convolutional neural network: a systematic review," *Int J Multimed Info Retr*, vol. 11, no. 3, pp. 315–331, Sep. 2022, doi: 10.1007/s13735-022-00242-9.

[32] P. Khatamino, İ. Cantürk, and L. Özyılmaz, "A deep learning-CNN based system for medical diagnosis: an application on Parkinson's disease handwriting drawings," in *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, IEEE, 2018, pp. 1–6.

[33] E. Irmak, "COVID-19 disease severity assessment using CNN model," *IET Image Processing*, vol. 15, no. 8, pp. 1814–1824, Jun. 2021, doi: 10.1049/ipr2.12153.

[34] A. A. Reshi *et al.*, "An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification," *Complexity*, vol. 2021, no. 1, p. 6621607, Jan. 2021, doi: 10.1155/2021/6621607.

[35] M. Bojarski *et al.*, "Visualbackprop: Efficient visualization of cnns for autonomous driving," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018, pp. 4701–4708.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[37] K. Zhang, P. Xu, and J. Zhang, "Explainable AI in deep reinforcement learning models: A shap method applied in power system emergency control," in *2020 IEEE 4th conference on energy internet and energy system integration (EI2)*, IEEE, 2020, pp. 711–716.

[38] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[41] Strubell E, Ganesh A, McCallum A., "Energy and policy considerations for modern deep learning research.," *InProceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 09, pp. 13693–13696, Apr. 2020.

[42] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," Mar. 20, 2015, *arXiv*: arXiv:1412.6572. doi: 10.48550/arXiv.1412.6572.

[43] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," Sep. 04, 2019, *arXiv*: arXiv:1706.06083. doi: 10.48550/arXiv.1706.06083.

[44] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," Oct. 17, 2018, *arXiv*: arXiv:1806.01261. doi: 10.48550/arXiv.1806.01261.

[45] Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K, Ghayvat H, "CNN variants for computer vision: History, architecture, application, challenges and future scope.," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.

[46] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351.

[48] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.