Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Advancements in Deep Learning Techniques for Image Recognition: A Comprehensive Review

Dr. Shashi Raj[1]
[1]Department of Computer Science and Engineering
Bakhtiyarpur College of Engineering, Bakhtiyarpur, Patna, Bihar, India
Email id- shashirajiitism@gmail.com

Ankita Sinha[2]
[2]Department of Computer Science and Engineering
Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Begusarai, Bihar, India
Email id- ankitasinha051@gmail.com

Reena Kumari[3]
[3]Department of Computer Science and Engineering
Bakhtiyarpur College of Engineering, Bakhtiyarpur, Patna, Bihar, India
Email id- reenanerist@gmail.com

Rajiv Kumar Ranjan[4]
[4]Department of Computer Science and Engineering
Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Begusarai, Bihar, India
Email id- rajivkr1234@gmail.com

Rohit Kumar[5]
[5]Department of Computer Science and Engineering
Rashtrakavi Ramdhari Singh Dinkar College of Engineering, Begusarai, Bihar, India
Email id- rohit05mgr@gmail.com

Abstract
Computer vision has been completely transformed by deep learning, especially when it comes to image identification applications. An extensive analysis of many deep learning architectures created for image recognition tasks is presented in this research study. This research examines the development of deep learning models, tracing their strengths, shortcomings, and performance on benchmark datasets from the earliest convolutional neural networks (CNNs) to the most recent state-of-the-art designs. The study also examines the crucial elements and design decisions that have aided in these architectures' success with picture recognition. It also covers the difficulties and potential avenues for further study in this dynamic and quickly developing topic.

Keywords— Artificial Intelligence, Deep Learning, Reinforcement Learning, Image Processing, Convolutional Neural Networks.

INTRODUCTION

One of the core tasks in computer vision is image identification, which is important for many applications such as surveillance, augmented reality, autonomous cars, and medical imaging. With the development of deep learning algorithms, the process of automatically identifying and classifying objects and scenes inside photographs has quickened significantly. In certain situations, deep learning architectures have outperformed human performance and conventional computer vision techniques in handling difficult picture identification tasks. An extensive analysis of the several deep learning architectures created for image recognition is presented in this research study. Deep learning has made tremendous strides in the last ten years, giving rise to a multitude

of potent models that have all brought new ideas and insights to the field. We set out on an expedition through the development of from the first convolutional neural networks (CNNs) to the most recent transformer-based models, of various designs. By proving that convolutional layers are a useful tool for feature extraction, the early CNNs, including LeNet-5, AlexNet, and VGGNet, set the stage for later developments. These ground-breaking models contributed significantly to the renaissance of neural networks and produced ground-breaking results in image recognition tasks, which advanced the field towards more complex structures. The vanishing gradient problem, which restricts the depth of conventional networks, is one of the main obstacles to training deep networks. Deep residual networks (ResNets) were developed as a solution to this problem. With the introduction of skip connections by ResNets, training ultra-deep networks became easier and information could go straight between layers. The paper's latter sections include a thorough investigation of the ResNet variations, each of which outperforms the others and achieves unmatched performance on a variety of image recognition tasks. Known as "GoogLeNet," inception designs significantly advanced the discipline by introducing the notion of "inception modules." To effectively capture multi-scale information, these modules use many filters of varying widths. The initial architecture was later improved upon by the Inception series, which included Inception v2, v3, and Inception-ResNet. These models gained notoriety for achieving great accuracy with comparatively fewer parameters. Dense connections between layers were introduced by DenseNet, a breakthrough in model construction. This invention made it possible to reuse features between layers, which led to a significant decrease in parameters without sacrificing model performance. By achieving cutting-edge results across several datasets, DenseNet established a new benchmark for parameter-efficient topologies. MobileNets have been developed in response to the increased need for delivering deep learning models on smartphones with limited resources. These models are ideal for mobile and embedded applications since they were designed to achieve great efficiency in terms of both memory footprint and computational needs. The pursuit of even more potent and efficient models gave rise to EfficientNet, which suggested a unique compound scaling technique to strike a balance between the depth, breadth, and resolution of the model. This family of models proved to be highly effective in a variety of resource-constrained circumstances, allowing for flexible deployment options. Additionally, the study looks at how transformer-based models alter picture recognition. Transformers were first created for tasks involving natural language processing. When they were modified for picture identification, the results were remarkable and new avenues for cross-modal learning research were created.

EARLY CONVOLUTIONAL NEURAL NETWORKS (CNNs)

The discipline of deep learning for image identification was greatly influenced by the early Convolutional Neural Networks (CNNs), which also set the foundation for the resurrection of neural networks in computer vision. Advances in a range of computer vision tasks were made possible by these groundbreaking models, which showed how well convolutional layers learned hierarchical information from pictures. Three key early CNN designs are covered in this section: LeNet-5, AlexNet, and VGGNet.

(i) LeNet-5- One of the earliest useful CNNs, LeNet-5 was created mainly for handwritten digit recognition and was originally presented by Yann LeCun et al. in 1998. Three convolutional layers, two subsampling (pooling) layers, and two fully connected layers made up its seven layers. Local patterns like edges and corners were learnt by the convolutional layers, while the subsampling layers decreased the spatial dimensions to allow translation invariance and lower the computing burden. LeNet-5 proved that CNNs are capable of handling image identification problems with their impressive performance on the MNIST dataset.

(ii) AlexNet - The 2012 proposal of AlexNet by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton was a major turning point in the development of deep learning. When this design was submitted to the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), it significantly reduced the error rate when compared to conventional computer vision methods. With eight layers total three fully linked and five convolutional AlexNet used a deeper architecture. It reduced the vanishing gradient issue and introduced non-linearity using the rectified linear unit (ReLU) activation function. In order to reduce overfitting and enhance generalisation, AlexNet also made use of dropout regularisation and data augmentation approaches. Deeper and more potent CNN architectures were later developed as a result of AlexNet's success.

(iii) VGGNet- Designed to investigate the effect of network depth on performance, VGGNet was first suggested in 2014 by the Visual Geometry Group at the University of Oxford. With 16 or 19 layers that included 2x2 max-

pooling layers and 3x3 convolutional filters, VGGNet's architecture was more homogeneous. VGGNet's depth enabled it to extract more intricate and abstract elements from pictures. VGGNet performed quite well on the ILSVRC 2014 dataset despite its depth, which makes it computationally demanding. This suggests that deepening the network might result in considerable accuracy increases.

DEEP RESIDUAL NETWORKS (RESNETS)

A novel class of deep learning architectures called Deep Residual Networks (ResNets) was created to solve the difficulties associated with training extremely deep neural networks. In their 2015 publication "Deep Residual Learning for Image Recognition," Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun introduced ResNets, which dramatically expanded the limits of model depth and demonstrated exceptional performance across a range of image recognition benchmarks.

(i) Motivation- The vanishing gradient problem, in which gradients transmitted down several layers diminish to almost zero, has made training very deep neural networks difficult and hindered early layers' ability to acquire meaningful representations. Traditional neural networks' depth was restricted by this problem, which also prevented them from performing as well as they might have. ResNets were created to solve this issue by employing skip connections or residual connections, which enabled even in extremely deep topologies for smooth gradient flow across the network.

(ii) Residual Blocks- Consisting of many convolutional layers, the residual block is the fundamental building component of ResNets. A residual block learns the residual mapping that is, the difference between the block's input and output instead of the intended mapping directly. Mathematically, the output of a residual block is calculated as follows given an input of x:

$$[\text{Output} = x + F(x) \]$$

where the mapping that the convolutional layers learnt is represented by F(x). By adding the original input x to the altered output, the residual connection essentially creates a "shortcut" path for the gradients to follow during backpropagation. This facilitates the optimisation and training of very deep networks by allowing the network to concentrate on learning the residual changes rather than learning the whole mapping.

(iii) Deep Residual Network design- A ResNet's whole design is made up of several residual blocks layered on top of one another. Usually, a series of residual blocks follows the first layers of the network, which carry out downsampling procedures (such as strided convolutions or pooling layers) to decrease spatial dimensions. Before the final classification layers, the network is subjected to upsampling procedures (such as transposed convolutions) to restore the spatial dimensions.

(iv) Variants- ResNets are available at several depths; some of the most well-known variations include ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The number represents the total number of layers in the network, comprising fully connected, batch normalisation, and convolutional layers. Although deeper variations, such ResNet-101 and ResNet-152, performed better on difficult tasks, their training also took longer and required more computer power.

(v) Impact- Deep learning and computer vision have greatly benefited from ResNets. They considerably outperformed shallower systems to win the ILSVRC 2015 image classification contest. Since then, ResNets have served as the basis for several more cutting-edge designs in a variety of computer vision applications, such as semantic segmentation, object identification, and picture creation.

TECHNIQUES/ TRANSFORMERS FOR IMAGE RECOGNITION

Transformers have showed a lot of promise for image identification and computer vision applications, while they were first presented for natural language processing tasks. Vision Transformers (ViT), often referred to as image recognition transformers, have drawn interest because of its capacity to manage long-range dependencies in pictures and outperform conventional convolutional neural networks (CNNs) in terms of performance. An important advance in computer vision is the use of transformers for tasks, which creates new opportunities for cross-modal and cross-domain learning.

(i) Vision Transformers (ViT)- Without using CNNs, Vision Transformers modify the transformer design to analyse pictures directly. Multi-layer perceptrons (MLPs) and self-attention processes are the fundamental elements of ViT. ViT uses self-attention to gather global contextual information, unlike CNNs that use fixed-sized kernels to process local areas. This enables it to effectively handle long-range relationships in pictures.

(ii) Self-Attention method- Vision Transformers' self-attention method enables every position to pay attention to every other position, capturing the contextual links between various visual elements. ViT's ability to comprehend the relationships and interactions between far-off visual areas, made possible by this global attention mechanism, increases its resistance to different image modifications and occlusions.

(iii) Patch Embeddings- An picture is separated into fixed-size, non-overlapping patches in order to convert it into the input format needed by the transformer. After that, each patch is linearly embedded to produce a series of vectors that are used as the transformer's input tokens.

(iv) Positional Encoding- To give the model positional information, positional encodings are added to the patch embeddings as transformers do not intrinsically encode the spatial information of the picture. The transformer can comprehend the spatial arrangement of patches and record their relative locations thanks to the positional encodings.

(v) Patch Embeddings- An picture is separated into fixed-size, non-overlapping patches in order to convert it into the input format needed by the transformer. After that, each patch is linearly embedded to produce a series of vectors that are used as the transformer's input tokens.

(vi) Positional Encoding- To give the model positional information, positional encodings are added to the patch embeddings as transformers do not intrinsically encode the spatial information of the picture. The transformer can comprehend the spatial arrangement of patches and record their relative locations thanks to the positional encodings.

(vii) Classification Head- Following the transformer layers' processing of the picture, the output is usually sent to a fully connected layer-based classification head that predicts the class labels or regression results.

(viii) Hybrid Techniques- It has also been investigated to use hybrid techniques that combine CNNs with transformers. For instance, some models combine CNN-based heads with transformers as a backbone to extract high-level characteristics and provide fine-grained predictions. These hybrid versions make an effort to combine the advantages of transformers with CNN efficiency.

CONCLUSION

In conclusion, significant progress has been made in computer vision and deep learning, which has completely changed how humans perceive and work with visual data. Researchers have consistently pushed the limits of image identification and computer vision tasks, starting with the development of convolutional neural networks and continuing with the emergence of cutting-edge designs like ResNets, Inception, DenseNet, MobileNets, and Vision Transformers. Key design decisions and elements that have contributed to the success and efficacy of many models during the trip include convolutional layers, activation functions, skip connections, attention processes, and more. Using benchmark datasets like ImageNet, CIFAR, and MNIST, these architectures have been tested and their strengths and weaknesses have been shown. Still, there are a number of issues to overcome, such as biases in the data and lack of interpretability to computational complexity and overfitting. These difficulties serve as stimulants for more study and creative thinking in the area. With fascinating developments in self-supervised learning, cross-modal comprehension, continuous learning, and explainable AI, the field of computer vision has a bright future ahead of it.

We anticipate seeing much more significant applications in fields like robotics, autonomous cars, healthcare, augmented reality, and surveillance as computer vision technology develops. AI's future will be shaped by its capacity to comprehend and analyse visual input, which will improve human-machine interactions and open up a wide range of positive uses for society.

REFERENCES

[1] Raschka, S.; Patterson, J.; Nolet, C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information **2020**, 11, 193.
[2] Barros, D.; Moura, J.; Freire, C.; Taleb, A.; Valentim, R.; Morais, P. Machine learning applied to retinal image processing for glaucoma detection: Review and perspective. BioMed. Eng. OnLine **2020**, 19, 20.
[3] Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A review of the application of machine learning in water quality evaluation. Eco-Environ. Health **2022**, 1, 107–116.
[4] Singh, V.; Chen, S.S.; Singhania, M.; Nanavati, B.; kumar kar, A.; Gupta, A. How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. Int. J. Inf. Manag. Data Insights **2022**, 2, 100094.
[5] Moscalu, M.; Moscalu, R.; Dascălu, C.G.; Țarcă, V.; Cojocaru, E.; Costin, I.M.; Țarcă, E.; Șerban, I.L. Histopathological Images Analysis and Predictive Modeling Implemented in Digital Pathology—Current Affairs and Perspectives. Diagnostics **2023**, 13, 2379.
[6] Wang, S.; Yang, D.M.; Rong, R.; Zhan, X.; Fujimoto, J.; Liu, H.; Minna, J.; Wistuba, I.I.; Xie, Y.; Xiao, G. Artificial Intelligence in Lung Cancer Pathology Image Analysis. Cancers **2019**, 11, 1673.
[7] van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. **2022**, 79, 102470.
[8] Prevedello, L.M.; Halabi, S.S.; Shih, G.; Wu, C.C.; Kohli, M.D.; Chokshi, F.H.; Erickson, B.J.; Kalpathy-Cramer, J.; Andriole, K.P.; Flanders, A.E. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiol. Artif. Intell. **2019**, 1, e180031.
[9] Smith, K.P.; Kirby, J.E. Image analysis and artificial intelligence in infectious disease diagnostics. Clin. Microbiol. Infect. **2020**, 26, 1318–1323.
[10] Wu, Q. Research on deep learning image processing technology of second-order partial differential equations. Neural Comput. Appl. **2023**, 35, 2183–2195.
[11] Jardim, S.; António, J.; Mora, C. Graphical Image Region Extraction with K-Means Clustering and Watershed. J. Imaging **2022**, 8, 163.
[12] Ying, C.; Huang, Z.; Ying, C. Accelerating the image processing by the optimization strategy for deep learning algorithm DBN. EURASIP J. Wirel. Commun. Netw. **2018**, 232, 232.
[13] Protopapadakis, E.; Voulodimos, A.; Doulamis, A.; Doulamis, N.; Stathaki, T. Automatic crack detection for tunnel inspection using deep learning and heuristic image post-processing. Appl. Intell. **2019**, 49, 2793–2806.
[14] Yong, B.; Wang, C.; Shen, J.; Li, F.; Yin, H.; Zhou, R. Automatic ventricular nuclear magnetic resonance image processing with deep learning. Multimed. Tools Appl. **2021**, 80, 34103–34119.
[15] Freeman, W.; Jones, T.; Pasztor, E. Example-based super-resolution. IEEE Comput. Graph. Appl. **2002**, 22, 56–65.
[16] Rodellar, J.; Alférez, S.; Acevedo, A.; Molina, A.; Merino, A. Image processing and machine learning in the morphological analysis of blood cells. Int. J. Lab. Hematol. **2018**, 40, 46–53.
[17] Kasinathan, T.; Uyyala, S.R. Machine learning ensemble with image processing for pest identification and classification in field crops. Neural Comput. Appl. **2021**, 33, 7491–7504.
[18] Yadav, P.; Gupta, N.; Sharma, P.K. A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods. Expert Syst. Appl. **2023**, 212, 118698.
[19] Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. Reinforcement learning coupled with finite element modeling for facial motion learning. Int. J. Multimed. Inf. Retr. **2022**, 11, 19–38.
[20] Zeng, Y.; Guo, Y.; Li, J. Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning. Neural Comput. Appl. **2022**, 34, 2691–2706.
[21] Pratap, A.; Sardana, N. Machine learning-based image processing in materials science and engineering: A review. Mater. Today Proc. **2022**, 62, 7341–7347.
[22] Mahesh, B. Machine Learning Algorithms A Review. Int. J. Sci. Res. **2020**, 9, 1–6.
[23] Singh, D.P.; Kaushik, B. Machine learning concepts and its applications for prediction of diseases based on drug behaviour: An extensive review. Chemom. Intell. Lab. Syst. **2022**, 229, 104637.

[24] Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the 4th International Conference on Learning Representations 2016, San Juan, Puerto Rico, 2–4 May 2016.
[25] Dworschak, F.; Dietze, S.; Wittmann, M.; Schleich, B.; Wartzack, S. Reinforcement Learning for Engineering Design Automation. Adv. Eng. Inform. **2022**, 52, 101612.
[27] Khan, T.; Tian, W.; Zhou, G.; Ilager, S.; Gong, M.; Buyya, R. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. J. Netw. Comput. Appl. **2022**, 204, 103405.
[28] Botvinick, M.; Ritter, S.; Wang, J.X.; Kurth-Nelson, Z.; Blundell, C.; Hassabis, D. Reinforcement Learning, Fast and Slow. Trends Cogn. Sci. **2019**, 23, 408–422.
[29] Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; Bowling, M. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science **2017**, 356, 508–513.
[30] ElDahshan, K.A.; Farouk, H.; Mofreh, E. Deep Reinforcement Learning based Video Games: A Review. In Proceedings of the 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 8–9 May 2022.
[31] Huawei Technologies Co., Ltd. Overview of Deep Learning. In Artificial Intelligence Technology; Springer: Singapore, 2023; Chapter 1–4; pp. 87–122.
[32] Le, N.; Rathour, V.S.; Yamazaki, K.; Luu, K.; Savvides, M. Deep reinforcement learning in computer vision: A comprehensive survey. Artif. Intell. Rev. **2022**, 55, 2733–2819.