IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.637

International Journal of Emerging Multidisciplinaries:

Computer Science and Artificial Intelligence

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (print): 2791-0164 ISSN (online): 2957-5036

AI-Enhanced Deep Learning Architectures for Image Recognition: A Comparative Analysis of CNN Models with Modern AI Integration

Lim Jia Ying [1] , Lu PengFei [1] , Low Hong Yi [1] , Chia Chen Yee [1] , Mshal Osama Gafar [1] , Mohammed Ali [1]

1. School of Computing and IT, Taylor's University (UWE Dual Awards Programme), Subang Jaya, Malaysia

Abstract

Convolutional Neural Networks (CNNs) have revolutionized image recognition by organizing features hierarchically and optimizing structures. In this paper, three popular CNN architectures (VGGNet, ResNet and GoogLeNet, also known as Inception) are compared in terms of their forward propagation, loss functions and back propagation. We delve deeper than that basic exploration and discuss the impact of current Artificial Intelligence (AI) developments such as Vision Transformers, self-supervised learning, and neural architecture search on the image recognition field. We analyze the computational efficiency, gradient flow, and scalability of each model, and illustrate that ResNet's residual connections provide the best of the three worlds of depth scalability, gradient stability and training efficiency. In addition, the emerging paradigms of AI like foundation models and multimodal learning are coming together with CNN-based solutions to shape the future generation of visual understanding systems. In this video, we will introduce you to the various neural network architectures that dominate the field of image recognition.In this video, we will introduce you to the various types of neural network architectures that dominate the field of image recognition: convolutional neural networks, deep learning networks, ResNet, VGGNet, GoogLeNet, Vision Transformers, self-supervised learning, and AI.

Keywords : *convolutional neural networks, image recognition, deep learning, ResNet, VGGNet, GoogLeNet, Vision Transformers, self-supervised learning, AI*

I. INTRODUCTION

The goal of image recognition is to let a computer automatically classify and recognize images. The most efficient way is the Convolutional Neural Network (CNN) that is used to classify images by the forward propagation [1]. During this process, the input image goes through a series of convolutional layers that are designed to capture the spatial structure of the image at various levels, including edges, textures, and shapes. The ReLU activation function is used to make the model learn more complex patterns, and pooling layers decrease the number of feature maps required, which helps to make the model more translation-invariant, more efficient, and more important for retaining key information [2]. High level features are then sent to fully connected or global average pooling layers to output a probability distribution for every category. Cross-entropy loss functions are used to compare the model's prediction of the true label and in backpropagation, the error is passed back up for parameter updates [2]. Different CNN architectures have been proposed to enhance the learning efficiency and accuracy over the years. VGGNet further extends the network hierarchy by stacking small convolutional kernels and pooling layers [3]; GoogLeNet extends the network hierarchy through multi-scale convolutional kernels that extract multi-level features in parallel [4]; ResNet extends the network hierarchy by incorporating residual connections to effectively mitigate the vanishing gradient problem, enabling much deeper networks [5]. These models have set the course for modern image recognition. In recent years, the Artificial Intelligence field launched paradigm shifting concepts, far beyond the traditional CNN structures. Recently Vision Transformers (ViTs) have been shown to perform competitive or better performance on image classification with training on large enough datasets, a feat that was previously thought to be exclusive to the NLP community [6]. In recent years, a number of self-supervised learning approaches have been developed to minimize the need for labeled data, which have learned to represent images using visual data without supervision [7] such as contrastive learning and masked image modelling. Moreover, the design of optimal network topologies with Neural Architecture Search (NAS) has been automated, resulting in networks such as EfficientNet that achieve better accuracy-efficiency trade-offs compared to hand-designed networks [8]. The paper begins with a detailed comparative study of VGGNet, ResNet and GoogLeNet architecture, loss function, backpropagation and behavior. It then puts these models into their respective perspectives within the new AI paradigm of foundation models, multimodal learning, and automated architecture design, providing a panoramic perspective of the evolving image recognition space.

II. MODEL ANALYSIS

A. VGGNet

1) Forward Pass:

VGGNet is a deep ConvNets network constructed on a simple and uniform architecture. It is made up of several 3×3 convolutional layers with ReLU activation and two 2×2 max-pooling layers, which progressively shrink the spatial dimensions and deepen the model [3]. The network consists of small 3×3 filters throughout the network which see fine details in the image, but keep the receptive field manageable. The typical VGG block comprises of two or three 3×3 convolutional layers that are followed by a max-pooling layer. The blocks are repeated for several times and flattening is applied after each block, followed by fully connected (FC) layers, prior to output classification layer. The last layer generates logits and the softmax activation [3] is applied to convert the logits into the probabilities of the different classes. VGG is easy to implement and extend due to its simple and uniform design. But its simplicity is accompanied by a drawback – it has a lot of parameters to set, particularly in the fully connected layers, which causes high computational and memory costs [3].

2) Loss Function:

VGGNet is built with the output layer having a Softmax activation and optimizes Categorical Cross-Entropy (CCE) loss for multi-class image classification. The true label y is one hot encoded, and the loss function used is the CCE loss: $L = -\sum y_k \log(\hat{y}_k)$ where logits z are converted to probabilities using Softmax. It's a gradual gradient which makes backpropagation more effective during training [2].

3) Backward Pass:

In backpropagation, the gradients of the loss function are propagated back through the fully connected layers, convolutional blocks and ReLU activations. The gradients of filters in each convolutional layer are calculated and error signals are propagated back to the previous layers via the chain rule [2]. The ReLU activation is important because it enables gradients to propagate efficiently. However, in very deep versions, since there are no shortcut connections, the gradients can also weaken when moving to the early layers of the VGG. Many later adaptations involve using Batch Normalization to boost the gradient stability [9].

B. ResNet

1) Forward Pass:

Residual Networks (ResNets) consist of a series of residual blocks which include multiple convolutional layers and a skip connection that feeds the input to the block to its output [5]. There are two kinds of skip connections: identity shortcuts (if the input and output dimensions are the same) and projection shortcuts (if the input and output dimensions are different; use 1×1 convolutions).

The residual block function is represented as $y = F(x, \{Wi\}) + x$, with x being the input, and y being the output, where $F(x, \{Wi\})$ is the residual mapping. The first layer in the network is a convolutional layer for extracting features, then followed by batch normalization for the stability of the training process, and finally max pooling for dimensionality reduction in the spatial dimension. Several residual blocks are cascaded and followed by global average pooling and a fully connected layer for classification [5].

Skip connections help reduce vanishing gradients and enable very deep networks to be trained efficiently by having gradients and information pass straight between layers. This was a key architectural breakthrough for showing that it is possible to effectively teach networks of more than 100 layers[17-19].

2) Loss Function:

ResNet uses Categorical Cross-Entropy (CCE) Loss for multi-class classification. The output logits are converted into probabilities by a Softmax function: $\hat{y}i = \exp(zi) / \sum \exp(zj)$. Lower the loss function $L = -\sum yi \log(\hat{y}i)$ is, the better the model is at predicting the true labels [2], [10].

3) Backward Pass:

Because the skip connections in ResNet enable the gradients to pass directly across the layers, very deep networks suffer from a less severe vanishing gradient problem. Even if the gradient from the residual mapping is small, the gradient from the shortcut still goes to the earlier layers, due to the "+1" term at each residual block [5]. This design provides a much easier way to optimize ResNet than a plain deep network of the same depth, and allows for efficient weight updates using gradient descent [5, 10].

C. Inception (GoogLeNet)

1) Forward Pass:

The idea behind GoogLeNet is to efficiently extract multiscale visual features by stacking multiple Inception modules [4]. Several branches are executed simultaneously in each module: 1×1 convolution branch, 1×1 to 3×3 convolution branch, 1×1 to 5×5 convolution branch, 3×3 max-pooling branch with 1×1 convolution. The 1×1 convolutions act as dimensionality reduction layers before the 3×3 and 5×5 convolutions that are more expensive, reducing the number of parameters and computation without sacrificing representational power [4] [11].

To further aid in gradient descent to earlier layers during training, GoogLeNet adds auxiliary classifiers that are connected to intermediate layers. They are supervised using an additional loss with a weight, and they are thrown out during inference [4].

2) Loss Function:

The Softmax activation over the last logits is used in GoogLeNet, and the Categorical Cross-Entropy (CCE) minimized. The total loss is used for training as Ltotal = Lmain + α·Laux1 + β·Laux2, where α and β are small values (around 0.3), and the auxiliary heads are eliminated at test time [4, 2].

3) Backward Pass:

Backpropagation begins from the output of the Softmax-CCE and is propagated down through the fully connected layer till the last Inception module. The upstream gradient at the concat node is sliced back to each branch, as the module concatenates the outputs of each branch together along the channel dimension. Each branch then backpropagates along its own sequence of convolutions and pooling operations [4]. In addition to this, auxiliary classifiers are used to add extra gradient paths which can help prevent vanishing gradients and stabilize learning in early layers [4, 10].

III. MODERN AI PARADIGMS IN IMAGE RECOGNITION

A. Vision Transformers (ViTs)

For more than a decade, CNNs have been the leading approach to image recognition, but Vision Transformers (ViTs) have become a powerful alternative. Dosovitskiy et al. [6] showed that image classification benchmarks can be successfully tackled with competitive performance by dividing an image into fixed-size patches, embedding them linearly, and feeding the sequence to a standard Transformer encoder. In contrast to convolutional operations, which suffer from locality constraints, the global attention component of the ViT can be used to model global dependencies between all image patches.

Unlike CNNs, however, ViTs generally need large-scale training sets (e.g., JFT-300M) to be more competitive than CNNs [6]. This limitation was overcome by hybrid architectures like DeiT (Data-efficient Image Transformers) and Swin Transformers, which introduce convolutional-like properties and hierarchical feature maps that deliver state-of-the-art performance while still being more data-efficient [12].

C. Attention and Memory Networks

The advent of self-supervised learning (SSL) has changed the way visual representations are learned. Some approaches like SimCLR, BYOL and DINO utilize pretext tasks or contrastive objectives to train models with meaningful features from unlabeled data [7]. It has been demonstrated that contrastive learning with data augmentations can generate visual representations comparable to supervised pretraining on ImageNet [7].

Masked image modeling, a technique inspired by the masked language modeling, has been applied in various ways such as Masked Autoencoders (MAE) [13]. These methods obscure random regions of an input image and have the model try to figure out how to recreate the missing pixels, learning valuable spatial and semantic information. These SSL techniques are especially important as they lower the requirement for costly human hand-annotation and make deep learning more scalable and available throughout domains[20-22].

D. Neural Architecture Search (NAS)

NAS: a system that automatically designs neural network architectures. NAS algorithms are used to search a defined configuration space to find optimal network configurations, instead of relying on human intuition. Found by NAS, efficientNet is a compound scaling method that scales network depth, width and resolution uniformly for improved accuracy and lower number of parameters and FLOPs than hand-designed networks [8].

The newer NAS methods include hardware aware optimization, which not only supports accurate discovery of architectures but also supports efficient use of discovered architectures on target deployment hardware, including mobile devices and edge computing hardware [14]. The synergy between architecture design and deployment efficiency is a crucial avenue for real-world AI systems.

E. Foundation Models and Multimodal Learning

They have recently been integrated as a unified language understanding model, commonly referred to as foundation models, like CLIP (Contrastive Language-Image Pre-training) [15]. CLIP is trained on 400 million image-text pairs with contrastive objectives and can instantaneously transfer to other visual classification tasks without task-specific fine-tuning. This multimodal approach shows that it is possible to achieve more generalizable and robust representations by combining visual and linguistic supervision.

Likewise, multimodal models such as Florence and Flamingo have shown the viability of few-shot learning and VQA for image recognition, indicating that the future of image recognition is not only controlled by a single modality of images, but also by multimodal intelligence [16]. These advancements represent a transformation in the paradigm, where CNN-like methods can be used as essential components in larger and more powerful AI systems.

IV. The critical evaluation and comparative analysis of the works.

In this section, the architecture design, computational efficiency and training performance of VGG, GoogLeNet and ResNet are analysed and compared, and also they are put into the context of the AI world[23-24].

VGG has a simple and uniform structure (stacked 3×3 convolutions), which is easy to implement and analyze, but a very large number of parameters (especially the fully connected layers), slows down the training speed and causes gradient attenuation without shortcut connections [3]. Even though it is not as impressive as other architectures, VGGNet's idea of a small, uniform filter has led to many other designs.

Using 1×1 convolutions for dimensionality reduction [4] and multi-scale feature extraction through parallel 1×1, 3×3 and 5×5 branches, GoogLeNet has the ability to extract many more features at less cost compared to the number of parameters. It has auxiliary classifiers to help with training. The difficulty to design and tune the architecture, however, makes it hard to create the direct gradient highways that residual connections offer [4] has yet to be achieved.

The analysis and comparison of the classical architectures, based on the result of this process, show that ResNet is the most suitable architecture to implement. The residual (skip) connections directly connect the input into the output of the block, enabling the gradients to follow a direct path, thus reducing the back-propagated signal's exponential decay [5]. This allows for the building of very large networks (such as ResNet-152) to converge and generalize on large datasets. ResNet is examined in this work with respect to three classical models that show the following advantages: depth scalability, gradient stability and training efficiency. In this work, ResNet is investigated in terms of the above three classical models: depth scalability, gradient stability and training efficiency.

In comparison to the recent strides in AI technology, ResNet's architectural advancements are still very relevant today. ResNet backbones have been extensively adopted in the field of Vision Transformers, object detection systems (such as Faster R-CNN), and multimodal models [6], [15]. For instance, the concept of skip connection was generalized in the architectures such as DenseNet and U-Net, which still have a profound impact [17].

New methods, like ViTs and foundation models, are now starting to do better than pure CNN models on large-scale benchmarks, however. For modern image recognition, hybrid solutions such as hybrid convolutional attention mechanism, self-supervised pretraining, and multimodal supervision are commonly used [12, 15]. These integrated approaches enhance the stability of the training process and optimize computational resources, further fostering sustainable and responsible deep learning practices.

V. CONCLUSION

This paper has explored in detail the three classic CNN architectures: VGGNet, ResNet and GoogLeNet, looking at their forward propagation, loss function and back propagation. In these classical models, ResNet is the one that has the best performance characteristics based on its residual connections, which tackle the vanishing gradient problem and contribute to the ability of training very deep networks.

In addition to this comparative study, we examined the impact of CNN architectures on the development of cutting-edge AI architectures and paradigms such as Vision Transformers, self-supervised learning, neural architecture search, and foundation models.In addition to this comparative study, we examined the impact of CNN architectures on the evolution of cutting-edge AI paradigms and architectures like Vision Transformers, self-supervised learning, neural architecture search, and foundation models. The paradigm of image recognition is moving towards systems that use attention mechanisms, multimodal supervision, and automated architecture design. With the continued advancement of the aforementioned technologies, VGGNet, ResNet, and GoogLeNet principles remain vital blocks in the expanding complex world of AI.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[2] F.-F. Li, J. Johnson, and S. Yeung, “CS231n: Convolutional Neural Networks for Visual Recognition — Lecture 3: Loss Functions and Optimization,” Stanford University, 2018.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, Sep. 2014.

[4] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–12.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[6] Khalil, M.I., Humayun, M., Jhanjhi, N.Z., Talib, M.N., Tabbakh, T.A. (2021). Multi-class Segmentation of Organ at Risk from Abdominal CT Images: A Deep Learning Approach. In: Peng, SL., Hsieh, SY., Gopalakrishnan, S., Duraisamy, B. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, vol 248. Springer, Singapore. https://doi.org/10.1007/978-981-16-3153-5_45

[7] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[8] Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing Drug Distribution Systems from Tampering Using Blockchain. *Electronics*, *11*(8), 1195. https://doi.org/10.3390/electronics11081195

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[11] Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing Drug Distribution Systems from Tampering Using Blockchain. *Electronics*, *11*(8), 1195. https://doi.org/10.3390/electronics11081195

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 448–456.

[13] Jhanjhi, N.Z. (2025). Investigating the Influence of Loss Functions on the Performance and Interpretability of Machine Learning Models. In: Pal, S., Rocha, Á. (eds) Proceedings of 4th International Conference on Mathematical Modeling and Computational Science. ICMMCS 2025. Lecture Notes in Networks and Systems, vol 1399. Springer, Cham. https://doi.org/10.1007/978-3-031-91005-0_43

[14] I. Azeem, "Loss functions in deep learning," Medium, Oct. 2023. [Online]. Available: https://medium.com/@ibtedaazeem/loss-functions-in-deep-learning-e4bd353ea08a

[15] Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abdelmaboud, A. (2022). A Trust-Based Model for Secure Routing against RPL Attacks in Internet of Things. *Sensors*, *22*(18), 7052. https://doi.org/10.3390/s22187052

[16] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.

[17] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.

[18] Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., & Humayun, M. (2023). A Deep Learning Approach for Atrial Fibrillation Classification Using Multi-Feature Time Series Data from ECG and PPG. *Diagnostics*, *13*(14), 2442. https://doi.org/10.3390/diagnostics13142442

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000–16009.

[20] B. Wu et al., "FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 10734–10742.

[21] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[22] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.

[24] Fatima-tuz-Zahra, N. Jhanjhi, S. N. Brohi, N. A. Malik and M. Humayun, "Proposing a Hybrid RPL Protocol for Rank and Wormhole Attack Mitigation using Machine Learning," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2020, pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257607.