IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.634

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

Deep CNN Architectures for Medical and General Image Recognition: Comparative Analysis with AI-Enhanced Perspectives

Wang Ruiting [1] , Khant Aung Chain [1] , Tao Jingchu [1] , Asser Tawfik [1] , Yuan Chengwei [1]

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

Abstract

In image recognition, using Convolutional Neural Networks (CNNs), a computer can be trained to recognize images by learning in a hierarchical way from pixel information. In this paper, three CNN landmarks, VGGNet, ResNet and GoogLeNet, are studied and analyzed in terms of forward propagation, categorical cross-entropy loss functions, and gradient flow in back-propagation. In the field of medical imaging, an example of ResNet implementation is used for classification of brain tumors in MRI images. We also delve into the ways in which current AI developments[17-19], such as Vision Transformers, self-supervised learning, neural architecture search, and multimodal foundation models, are reshaping image recognition from traditional CNN methods.

Keywords—convolutional neural networks, image recognition, ResNet, VGGNet, GoogLeNet, medical imaging, brain tumor classification, Vision Transformers, deep learning

I. INTRODUCTION

Image recognition is a type of computer pattern recognition that uses computer vision to recognize images by learning the pixel patterns of the image [1]. Images are first encoded into numerical pixel values at the input layer (e.g. a 50×50 RGB image would have 7,500 input features). Hidden convolutional layers then learn increasingly complex features in a hierarchical fashion—low-level layers are good at identifying

edges and colors, mid-level layers are good at matching shapes and textures, and high-level layers are good at matching abstract concepts [2, 3].

Inputs into each neuron are multiplied by a learned weight, and activation functions (ReLU, sigmoid, tanh) are applied to impart nonlinearity. The output layer, usually consisting of one neuron for each class, applies softmax to produce outputs that are probabilities which sum to 1 and can be used as the prediction for the class label [1]. Categorical cross-entropy loss is used as a measure of prediction error, and backpropagation is a method of tuning the weights iteratively, using gradient descent [1, 4].

There are three important architectures which address these processes in different ways. The VGGNet uses uniform 3×3 stacked convolutions for deep sequential processing [5]. ResNet: Extremely deep networks with residual blocks with skip connections [6]. In order to extract features efficiently, GoogLeNet adopts parallel multi-scale Inception modules [7]. Both have unique advantages in terms of gradient stability, speed, and scalability in terms of depth[18-20].

Emerging paradigms in modern AI have emerged, including Vision Transformers with self-attention for the global understanding of images [8], self-supervised learning for minimizing the need for labeled data [9], neural architecture search for automated model design [10] and foundation models such as CLIP for multimodal visual-linguistic reasoning [11]. The paper reviews the classical architectures, showcases an implementation of ResNet for brain tumor classification and puts these in the wider context of the AI landscape.

II. MODEL ANALYSIS

A. VGGNet

1) Forward Pass:

VGGNet is a very simple and uniform architecture of 3×3 convolutional layers with ReLU activation functions and 2×2 max-pooling with stride 2, repeated periodically. This design is used to increase the depth progressively while maintaining fine spatial detail [5]. The maps are flattened, and then passed through fully connected layers, where class probabilities are produced by the last softmax layer. The small filter approach is consistent, which results in an efficient complex pattern capture with manageable parameters [5].

2) Loss Function:

VGGNet is based on Categorical Cross-Entropy loss: $L = -\sum y_i \log(\hat{y}_i)$, where y_i is the one-hot true label and $\hat{y}_i$ is the softmax prediction. Larger losses mean poorer predictions, and smaller losses mean that the learning is good. In gradient-based optimization [4, 5] the loss is the main feedback signal.

3) Backward Pass:

VGGNet achieves gradient stability by using ReLU activations (which have a derivative of 1 when they are active) and small convolutional kernel sizes, which means that gradients have a predictable magnitude, as well as by having the same pattern in each layer, which means that they have a smooth flow [5]. But the sequential design without shortcut mechanisms is still prone to gradient vanishing with deep versions [5] [6].

B. ResNet

1) Forward Pass:

ResNet's residual blocks include identity skip connections, with the output being $y = F(x) + x$. The network isn't learning the transformations themselves, but instead the residual mappings, and the shortcuts carry lower level information [6]. Deeper variants are enhanced by bottleneck structures to increase computational efficiency[21]. Final predictions are obtained by using a fully connected softmax layer with global average pooling [6].

2) Loss Function:

ResNet uses the same loss as CCE. The gradient with respect to the output layer, $\partial L/\partial z_k = \hat{y}_k - y_k$, directly gives an optimization signal that can be used for adjusting weights using backpropagation [4] and [6] .

3) Backward Pass:

Skip connections divide up the flow of gradients in two directions, one through the residual branch and one through the identity branch. Another way of insuring that gradients reach earlier layers without reduction is to use the identity path, which has the property $\partial y/\partial x = \partial F/\partial x + I$ (I = identity matrix) [6]. This allows you to train networks with more than 150 layers.

C. GoogLeNet (Inception)

1) Forward Pass:

The Inception modules support parallel branches (1×1, 3×3 convolutions and pooling), with outputs concatenated in the depth dimension, to enable multi-scale feature representation [7]. The number of parameters is kept under control by using 1×1 convolutions to reduce the dimensionality. The network processes several modules, each making progressively more sophisticated feature extraction [7].

2) Loss Function:

Augmenting the CCE loss with auxiliary classifiers at intermediate layers is used with GoogLeNet. There are additional gradient signals for stabilized learning from the total training loss $L_{total} = L_{main} + \alpha L_{aux1} + \beta L_{aux2}$ [7].

3) Backward Pass:

Each branch of the Inception has its own gradient which flow independently and combine at concatenation points, which means that many gradient pathways were created which minimizes the risk of vanishing. Auxiliary classifiers bypass intermediate layers and add gradients directly to them, while 1×1 convolutions reduce the number of parameters to control the magnitude of the gradients [7].

III. MODERN AI PARADIGMS

A. Vision Transformers

ViTs represent images as a sequence of patches and employ Transformer encoders to model global dependencies by self-attention [8]. The Swin Transformers are a combination of hierarchical features with shifted windows for efficient state-of-the-art performance [12]. Hybrid CNN-Transformer architectures make use of both local and global processing power.

B. Self-Supervised Learning

The contrastive methods (SimCLR, BYOL) and masked image modeling (MAE) do not require labeled data to learn visual representations [9], [13]. This is a significant benefit for medical imaging with the need to hire an expert annotator, who is both very expensive and not readily available[15-17].

C. Deep Learning for Computer Vision

Architecture discovery is automated and EfficientNet performs better on scaling [10]. In addition, foundation models such as CLIP allow few-shot classification from multimodal pretraining [11] and Flamingo further extends this to few-shot visual reasoning [14].

IV. IMPLEMENTING THE RESNET IN THE BRAIN TUMOR CLASSIFICATION.

A. Designed for a specific type of data.C. Designed for architecture.

A model based on ResNet was implemented for classification of brain tumors from MRI images. The architecture features residual blocks, ReLU activations and batch normalization. Images (128 x 128 x 3) are fed into the convolutional blocks, which have 32, 64 and 128 filters, and then undergo Global Average Pooling, Dense(128) and softmax classification blocks [6]. Data augmentation (rotation, shifting, flipping) and 80/20 train-validation split was also used.

An Adam optimizer (with learning rate 0.001) was employed [15]. The training accuracy increases from 53% to 88% after 20 epochs and the validation accuracy reaches its peak at epoch 12, with a value of 73.30%. The model had a loss value of 4.77 and a test accuracy of 38.6%, which shows that the model

does not perform well in generalizing its knowledge to unseen test data, especially in small medical imaging datasets. The disparity between the training and test results indicate that more data augmentation, progressive unfreezing or self-supervised pre-training is required for better generalization [6, 9].

V. CRITICAL EVALUATION

Although VGGNet is simple, it is easy to implement and has a good gradient flow analysis, but it has a high computation cost and a large gradient drop in deep architecture [5]. In order to exploit multi-scale features with efficient parallel architecture and support gradients [7] GoogLeNet has been designed. The skip connections used in ResNet are the most powerful gradient preservation, which allow the convergence of very deep networks and the stability of their training in addition to their good transfer learning performance [6].

ResNet was chosen for its gradient stability and scalability in depth for the brain tumor classification task. When implemented, it was found that small datasets and visual similarity between classes poses some problems in medical imaging. Self-supervised pretraining [9] and domain-specific foundation models have been proposed as potential directions for further advancements in these difficult medical imaging tasks, as is hybrid CNN-Transformer architectures [8], [12].

VI. CONCLUSION

In this paper, VGGNet, ResNet and GoogLeNet was compared and contrasted with regards to how they managed the flow of the gradient. Brain tumor classification with ResNet showed decent training convergence and pointed out the difficulties of generalization in medical imaging. Building on these core principles, the following modern AI paradigms are leveraging these classical CNN architectures in various ways, with transformative implications for general and medical image recognition: Vision Transformers, self-supervised learning, NAS, and multimodal foundation models.

REFERENCES

[1] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” Neural Comput., vol. 29, no. 9, pp. 2352–2449, 2017.

[2] R. H. Abiyev and M. K. S. Ma’aitah, “Deep convolutional neural networks for brain tumor classification,” Cogn. Syst. Res., vol. 72, pp. 68–79, 2018.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NeurIPS, 2012, pp. 1097–1105.

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

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.

[6] Ghosh, S., Singh, A., Kavita, Jhanjhi, N. Z., Masud, M., & Aljahdali, S. (2022). SVM and KNN based CNN architectures for plant classification. *Computers, Materials & Continua, 72*(1), 1927–1945. https://doi.org/10.32604/cmc.2022.023414

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.

[8] Zaman, S. K. u., Jehangiri, A. I., Maqsood, T., Umar, A. I., Khan, M. A., Jhanjhi, N. Z., Shorfuzzaman, M., & Masud, M. (2022). COME-UP: Computation Offloading in Mobile Edge Computing with LSTM Based User Direction Prediction. *Applied Sciences*, *12*(7), 3312. https://doi.org/10.3390/app12073312

[9] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE CVPR, 2015, pp. 1–12.

[10] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.

[11] T. Chen et al., “A simple framework for contrastive learning of visual representations,” in Proc. ICML, 2020, pp. 1597–1607.

[12] 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

[13] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105–6114.

[14] 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.

[15] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021, pp. 8748–8763.

[16] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. ICCV, 2021, pp. 10012–10022.

[17] 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

[18] K. He et al., “Masked autoencoders are scalable vision learners,” in Proc. CVPR, 2022, pp. 16000–16009.

[19] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” in Proc. NeurIPS, 2022.

[20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.

[21] N. Srivastava et al., “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929–1958, 2014.