IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.635

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-Driven Deep Learning for Flower Image Recognition: Comparative CNN Analysis and ResNet50 Transfer Learning Implementation

Juston Tan Yi Xian [1] , Chua Li Ling [1] , Gam Chui Ern [1] , Goo Yun Hai [1] , Rebecca Law Wen Qi [1]

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

Abstract

Convolutional Neural Networks (CNNs) have transformed the field of computer vision, enabling computers to recognize and classify visual data with high accuracy. This paper compares the three well known CNN architectures namely VGGNet, ResNet and Inception (GoogLeNet), specifically while looking at their forward propagation, loss functions, and gradient flow mechanisms. We propose a ResNet50 based transfer learning system for flower species classification on the TensorFlow Flowers dataset with the test accuracy of 90.33%. In addition to classical methods, we examine the application of contemporary AI methods such as Vision Transformers, self-supervised learning, neural architecture search, and multimodal foundation models, which are enhancing the capabilities of image recognition and complementing the CNN based methods.

I. INTRODUCTION

The image recognition is one of the most important applications of Artificial Intelligence, which involves recognizing and categorizing visual data [1]. Traditional image recognition systems used to depend on manually specified characteristics like colour, shape or texture, which were slow to process and could not handle lighting conditions, background or angles of the camera [2]. Deep learning models are able to

automatically learn key features from the analysis of large datasets, among the most successful models are Convolutional Neural Networks (CNNs) [3].

CNNs are a series of interconnected layers that process images in a hierarchy, with early layers detecting simple features such as edges, and later layers detecting more complex patterns and shapes [4]. In the forward pass, the data passes through the convolutional layers, pooling layers, and fully connected layers, and class predictions are output. The prediction loss function is categorical cross-entropy loss and weight modification is done using back-propagation technique [2, 5].

Several CNN architectures have been developed over the years to be more efficient and accurate. VGGNet uses stacked small convolutions to extract deep features [3], GoogLeNet uses parallel multi-scale Inception modules [6] and ResNet introduces skip connections to achieve training of very deep networks by overcoming the problem of vanishing gradients [5]. These models are the basis of current image recognition systems.

In more recent years, new AI paradigms have branched off from the conventional CNN designs. Self-attention mechanisms are used to process patches of an image in Vision Transformers (ViTs) to get competitive performance [7]. The use of a self-supervised learning framework decreases reliance on labeled data [8] and Neural Architecture Search (NAS) automates the search for optimal architectures [9]. The foundation models, such as CLIP [10] serve as a multimodal pretraining approach for connecting visual and textual understanding. In this paper, we investigate the three classic CNN architectures and then implement ResNet50 network on flower classification, and finally put these networks in the context of today's modern AI[22-23].

II. MODEL ANALYSIS

A. Forward Pass

In a forward pass, the data flows from the input in a sequential manner through the hidden layers to the output in CNN. Convolutional layers are used to generate feature maps that are sensitive to certain visual aspects of the image, like textures, corners and edges [4]. Feature maps grow increasingly abstract as they become deeper in the data base, representing shapes, parts of objects, and semantic structures [1]. The feature maps of each dimension after convolution and pooling are flattened into a single dimension, and then processed by fully connected layers. After passing through the logits of the last layer, they pass through a softmax function and form a normalized probability distribution [6].

1) VGGNet:

VGGNet is a deep and strictly sequential network, with a number of stacked 3×3 layers. This allows for progressive depth while retaining fine spatiotemporal information and learning abstract representations. They are built sequentially, increasing in interpretability and structural uniformity. In deeper versions however, because of the high number of parameters, high computation cost and vanishing gradient problem arise [3],[1].

2) ResNet:

ResNet adds “residual” (or “skip”) connections that allow the model to skip any number of layers by simply connecting the input to the output. The learned residual mapping is collected as the residual block function $y = F(x, \{Wi\}) + x$. This helps prevent any noisy gradients from occurring during backpropagation allowing the training of networks of hundreds of layers that will learn well for long periods without changing its learning rate [5, 6].

3) Inception (GoogLeNet):

Each of the inception modules is in parallel to perform multiple convolutional operations with kernels of size (1×1, 3×3, 5×5), and their outputs are concatenated along the depth dimension. The 1×1 convolutions are used to lower the computational cost before the larger convolutions to capture detailed features and global features at the same time [6]. Finally, instead of fully connected layers, GoogLeNet ends with global average pooling to decrease the number of parameters and risk of overfitting [6].

B. Loss Function

For multi-class classification, all three architectures uses Categorical Cross-Entropy (CCE) loss. The loss is defined as $L = -\sum yi \log(\hat{y}i)$, with yi being the one-hot encoded true label, and $\hat{y}i$ being the prediction probability from the softmax output [2] [11]. The smaller the loss, the better the performance of the model. GoogLeNet also adds auxiliary classifiers that have weighted losses ($Ltotal = Lmain + 0.3(Laux1 + Laux2)$) to enhance the flow of gradients when training the network [6].

A. Fw. Backward Pass, Gradient Flow

In the backward pass, the chain rule is applied to calculate the change in the loss with respect to each weight. One of the main challenges is the vanishing gradient problem [1] that arises in deep networks as the gradients fade away exponentially. VGGNet's depth is sensitive, where the multiplication of small derivatives through many layers in the network reduces the gradient signal to early layers [3], [6].

In ResNet, it is solved using skip connections which are shortcuts for the gradients. Instead of learning a complete mapping H(x), ResNet learns a residual mapping F(x) = H(x) – x, letting gradients pass directly from the latter layers to the former. This avoids the problem of vanishing gradients and makes it possible to train very deep networks [5;6].

However, Inception works around vanishing gradients by using a multi-branch architecture to apportion gradient signals across multiple computational routes, minimizing the chances that any one route would be too weak to propagate information. Auxiliary classifiers give extra gradient information to the intermediate layers to further stabilize learning in the early layers [6].

III. The new paradigm of Artificial Intelligence for image recognition.

A. Vision Transformers (ViTs)

The idea of Vision Transformers is to break down images into fixed-size patches, linearly embed them, and pass the sequence to the Transformer encoders to process with self-attention mechanisms [7]. The ViTs model global dependencies across all the patches of the image, a locality constraint that is overcome by convolutions. However, because of the lack of translational equivariance, they do not outperform CNNs in a large scale pre-training setting. Hybrid architectures like Swin Transformers combine hierarchical features with shifted-window attention for state-of-the-art performance with practical data requirements [12].

B. Self-Supervised Learning

Self-supervised learning (SSL) approaches like SimCLR and BYOL have been developed to learn useful visual representations from unlabeled data via contrastive objectives [8]. Masked Autoencoders (MAE) randomly mask patches of the image and restore the missing parts, capturing rich spatial features [13]. The impact of SSL will be profound in scaling up and making deep learning more applicable.

C. Neural Architecture Search

NAS automatically recognizes the best architectures. EfficientNet, found by NAS, added compound scaling of depth, width and resolution to produce better accuracy with the fewer parameters [9]. The hardware-aware NAS approaches guarantee that discovered architectures are efficient on mobile and edge platforms [14].

D. Foundation Models

CLIP is trained on 400 million image-text pairs, which allows it to perform zero-shot visual classification without the need for task-specific fine-tuning [10]. Few-shot learning also brings multimodal capabilities to image recognition, which suggests that future image recognition will combine visual, linguistic and contextual intelligence, e.g. through the model Flamingo [15].

IV. RESNET50 IMPLEMENTATION

A. Dataset and Preparation

There are 3,600 images of 5 species in the TensorFlow Flowers dataset: daisy, dandelion, roses, sunflowers, and tulips. The images were preprocessed to be 128×128 sized and split into training and testing sets (80/20) using tf.data pipelines, then normalized, shuffled, and batching. The class labels were one hot encoded to fit the CCE loss function [11].

B. Model Architecture

This feature extraction backbone was ResNet50, pre-trained on ImageNet, without the include_top argument [5]. The base layers were all set to be trainable for fine-tuning. A Global Average Pooling layer was used to compress the 7×7×2048 feature maps, followed by a Dense(512, ReLU) layer for non-linear feature combination, a Dropout(0.5) layer for regularization, and Dense(5, Softmax) layer for five-class classification.

C. Training and Optimization

Optimization was performed with the Adam optimizer using a learning rate of 1e-5 with momentum and RMSProp [16-18]. The best model on validation loss has been saved by ModelCheckpoint. The loss decreased quickly at the start and then levelled off for the training performed over 100 epochs. The highest validation accuracy of 93.32% and validation loss of 0.2381 were reached at epoch 24, after which, there was some sign of overfitting[19-21].

D. Evaluation Results

The model has a test accuracy of 90.33% and test loss of 0.2855. Per-class F1-scores were: dandelion (0.97), daisy (0.96), sunflower (0.93), tulip (0.88), and rose (0.81). The lower scores for roses and tulips are due to the fact that they are similar in visual terms. Visualization of the softmax indicated that correctly classified images displayed strong bars of probabilities with a sharp peak denoting high model confidence [5] [11].

V. CRITICAL EVALUATION

With the three architectures, VGGNet is simple, which helps to understand, but has much of parameters, but does not have the gradient management mechanism, which is difficult to do deep training [3]. It is designed to be more efficient and extract multi-scale features, as in GoogLeNet where the inception modules are added [6]. ResNet's skip connections offer the best gradient flow and allow for very deep networks to be trained without the network converging [5].

The results showed that ResNet50 has superior theoretical benefits, and its practical implementation, based on transfer learning, achieved an accuracy of 90.33% in the classification of flowers. ResNet backbones are still popular in Vision Transformers, object detection, and also in multimodal models [7, 10]. New hybrid methods that incorporate convolutional extraction methods with attention and/or self-supervised pretraining, however, are becoming the norm for the state-of-the-art [7, 12, 13].

VI. CONCLUSION

This paper has theoretically analyzed VGGNet, ResNet, GoogLeNet architectures and showed that ResNet50 is the best among them by using practical flower classification with an accuracy of 90.33%. The core principles of classical CNN architectures remain vital in this rapidly evolving AI landscape, while the new AI paradigms such as Vision Transformers, self-supervised learning, and foundation models are continuing to expand the scope of classic CNNs.

REFERENCES

[1] L. Alzubaidi et al., “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” J. Big Data, vol. 8, no. 1, pp. 1–74, 2021.

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

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, pp. 1097–1105.

[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] 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

[7] Saeed, S., Abdullah, A., Jhanjhi, N.Z. *et al.* New techniques for efficiently k-NN algorithm for brain tumor detection. *Multimed Tools Appl* **81**, 18595–18616 (2022). https://doi.org/10.1007/s11042-022-12271-x

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

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

[10] 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

[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] 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.

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

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

[15] 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

[16] F.-F. Li, J. Johnson, and S. Yeung, "CS231n: Convolutional Neural Networks for Visual Recognition," Stanford University, 2018.

[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] 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.

[19] 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.

[20] B. Wu et al., "FBNet: Hardware-aware efficient ConvNet design via differentiable NAS," in Proc. IEEE/CVF Conf. CVPR, 2019, pp. 10734–10742.

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

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[23] 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