IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.636

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-Augmented Convolutional Neural Networks for Image Recognition: Architectural Analysis and ResNet50 Implementation with Modern Deep Learning Perspectives

Nicolas Lee [1] , William Melvin Sukamto [1] , Jovin Maurelio [1] , Jovan Maurelio [1]

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

Abstract

The Convolutional Neural Network (CNN) is now the backbone of image recognition technology, powering computers to make accurate classifications and understand visual content. In this paper, we do a detailed analysis of the architecture of three popular CNN models namely ResNet, GoogLeNet (Inception v1) and VGGNet, and discuss the forward propagation, loss functions and back propagation mechanisms. We also explain in detail a practical implementation of ResNet50 on the CIFAR-10 dataset, which involves data preparation, building of the model, optimization of training using the Adam optimizer, and evaluation. In addition to this classic analysis, we discuss the latest developments in artificial intelligence paradigms such as Vision Transformers (ViTs), self-supervised learning, neural architecture search (NAS), and multimodal foundation models, which are transforming the image recognition field and supplementing what was already achieved by CNN models.

Keywords : convolutional neural networks, image recognition, ResNet50, transfer learning, VGGNet, GoogLeNet, Vision Transformers, self-supervised learning, deep learning

I. INTRODUCTION

As humans, we can easily identify and distinguish images without much effort. For computer, the task of image recognition and understanding is much more complicated, because computer only "sees" the images

as numbers of pixels [1]. In the era of image recognition technology, images can now be analyzed, classified, and interpreted more efficiently by a computer: product identification, defect detection, facial recognition – and more – can be done in a very short time and with great accuracy [2].

Image recognition uses machine learning algorithms, particularly Convolutional Neural Networks (CNNs), to recognize and categorize features and patterns in images. In training, pictures are fed into a series of layers of a CNN, and the sequential patterns are learnt by identifying edges, shapes and textures in the images [1]. CNN is applied to a forward pass, with the input data passing through convolutional layers, pooling layers, and fully connected layers, and then producing class predictions. A loss function compares the prediction with the true label and backpropagation updates the weights to better predict in the future [1,3].

As time has passed, several CNN models have been proposed to boost learning efficiency and accuracy. VGGNet adds the advantage of stacking small convolutional kernels to deepen the network [4]; GoogLeNet adds the advantage of using multi-scale convolutional kernels via the Inception modules to extract features in parallel [5]; ResNet adds the advantage of residual connections to address the vanishing gradient problem and allow for networks that are significantly deeper [6]. The architectures have been all of them responsible for setting the course of modern image recognition.

The past few years have seen significant paradigm-shifting advances in the realm of AI that go beyond the standard CNN architecture. The Vision Transformers (ViTs) have shown that attention-based models can perform competitive or better classification on images [7]. On the one hand, the use of self-supervised learning methods has significantly lowered the reliance on labeled data [8], whereas on the other hand, the optimization of network architecture, called Neural Architecture Search (NAS), automated network structure optimization [9]. Previously, researchers have enabled the integration of visual and textual understanding by using multimodal pretraining in the context of foundation models like CLIP [10].

This paper first presents a detailed comparison between ResNet, GoogLeNet and VGGNet. It then introduces a practical implementation of ResNet50 based on transfer learning on the CIFAR-10 dataset, and finally discusses the role of these models in the context of the growing AI ecosystem of ‘foundation models, multimodal learning and automated architecture design.

II. MODEL ANALYSIS

A. Residual Networks (ResNet)

1) Forward Pass:

The input image is fed through a series of convolutional layers in ResNet to learn hierarchical features. The main element of architecture is the residual block which is a set of layered and skip connections which

allow the network to learn the difference between the input and output, instead of learning the entire transformation directly [6]. In a residual block, the input x goes through the convolutional layers to generate F(x). In addition to the use of F(x) only, the original input x is added in the so-called skip connection, creating y = F(x) + x which is fed into the subsequent block.

The feature extraction process output feature maps are globally averaged and sent to a fully connected layer, which is used to generate classification logits (raw prediction scores for each class). Skip connections enable the flow of gradients and information directly between layers which helps avoid vanishing gradients and make it easier to train very deep networks [6].

2) Loss Function:

For multi-class classification, ResNet adopts a standard loss function, which is called categorical cross-entropy (CCE) loss. Logits are then passed into softmax to convert them to probabilities after the forward pass. The CCE loss is mathematically formulated as $L = -\sum y_i \log(\hat{y}_i)$, where y_i is the truth label (1 if it is the right label, 0 otherwise), $\hat{y}_i$ is the probability value predicted by the softmax output [3] [11]. If the predicted probability for the correct class is low, the loss will grow as the value of the logarithm will be larger, only the correct class contributes to loss.

3) Backward Pass:

The vanishing gradient problem [6] is a common phenomenon in a traditional deep CNN as the network becomes deeper, the gradients during the back-propagation become smaller. To solve this, ResNet makes use of skip connections, which directly connect the outputs to the inputs of residual blocks. In backpropagation, gradients pass through the shortcut path as well as through the primary convolutional path. This allows for training of the model with strong learning signals for the early layers as well, which not only results in easier training but also quicker convergence and allows for much deeper architectures to be supported [6].

B. GoogLeNet (Inception v1)

1) Forward Pass:

GoogLeNet does not just use one filter size per layer but rather multiple filter sizes (1×1, 3×3, 5×5) and pooling operations at the same time [5]. The output of each branch is fused vertically to create a single output feature map. This allows model to capture fine-grained details with 1×1 convs and coarse-grained with 5x5 convs.

The 1×1 convolutions play an essential role to act like dimensionality reduction layers, which have all the input channels, but also reduce the depth, so that the following larger convolutions have a significantly lower computational cost [5, 12]. To reduce the number of parameters and the risks of overfitting,

GoogLeNet applies Global Average Pooling instead of fully connected layers to compute the mean value of each feature map in order to obtain a compact summary of the feature map contents.

2) Loss Function:

GoogLeNet also adopts categorical cross-entropy loss as the training loss. But it also includes additional classifiers that have their own loss functions. The total loss is calculated as Ltotal = Lmain + 0.3(Laux1 + Laux2) [5] with a small weighting factor 0.3. By using auxiliary losses, gradient propagation is enabled more effectively when backpropagating through the network. During testing, auxiliary classifiers are discarded, and a model is only based upon the main softmax output [5].

3) Backward Pass:

GoogLeNet has 22 layers and gradients tend to decrease as we go backward. The auxiliary classifiers serve as gradients checkpoints, sending gradients back to earlier layers and improving training efficiency while alleviating the vanishing gradient problem [5]. Also, each gradient is calculated in isolation from the others in the convolution paths of each Inception module, which promotes the same learning across the network [5], [3].

C. VGGNet (VGG-16)

1) Forward Pass:

The basic design goal of VGGNet is to develop a deep network with a simple and uniform architecture, and the key is to use multiple small convolutional filters (3×3). VGG-16 has 13 convolutional layers and 3 fully connected layers. In the forward pass, the normalized input image (224×224×3) is passed through a series of convolutional layers with ReLU activations, followed by 2×2 max-pooling layers, which help to reduce the dimensionality of the image in a way that preserves the important information. The output is flattened and fully connected layers are applied to it, and the final softmax layer outputs class probabilities from the logits [4].

2) Loss Function:

The loss function of VGGNet is the same as that of ResNet and GoogLeNet: $L = -\sum yi \log(\hat{y}i)$. For instance, if the classification of an image is {Elephant, Tiger, Lion, Whale} and the true class is Lion, while the probabilities of the four categories are [0.05, 0.15, 0.75, 0.05] then the value of L is about $-\log(0.75) \approx 0.288$. The larger the loss, the further away from the actual label the prediction is made [3, 4].

3) Backward Pass:

VGGNet suffers from the vanishing gradient problem, which is caused by its deep sequential architecture, because the gradients decrease as they go back to earlier layers. VGGNet does not have a specific architectural mechanism to overcome the problem as ResNet or GoogLeNet does, apart from the ReLU

activation which provides some gradient preservation [4]. This restriction finally inspired the creation of ResNet, which was explicitly proposed to alleviate gradient degradation in deep networks [6].

III. THE MODERN WORLD OF IMAGE RECOGNITION – MODERN AI PARADIGMS IN IMAGE RECOGNITION

A. Candidate Augmented Transformers (CATs)

Vision Transformers (ViTs) have shown great promise as an alternative to CNNs, particularly for their ability to process images with high accuracy and efficiency.Over the last decade, CNNs have been the leading approach for image recognition, but recently, Vision Transformers (ViTs) have gained traction as a powerful alternative. Dosovitskiy et al. [7] showed that the linear embedding of a fixed-size patch and processing the sequence with a standard Transformer encoder provides competitive results on image classification datasets. Unlike convolutional operations, which have a locality constraint, ViTs are powerful enough to capture global dependencies across all image patches through self-attention.

The main drawback of the ViTs is that they are harder to outperform CNNs, generally because they need large-scale pretraining datasets to do so, due to the absence of the inductive bias of translation equivariance [7]. To improve practicality of the data requirements, hybrid architectures are developed, for instance, Swin Transformers, which combine shifted-window attention with hierarchical feature maps, yielding world-leading performance [13]. The trend is towards integrating convolutional and attention mechanisms in future image recognition systems.

B. Knowledge Graphs: Structure and Applications

Self-supervised learning (SSL) has revolutionised the learning of visual representations. For instance, models can be trained using a task like SimCLR or BYOL to extract meaningful features from unlabeled data [8]. It was demonstrated by Chen et al. [8] that representations using contrastive learning with data augmentation strategies can match supervised ImageNet representations.

Masked image modeling, which is similar to BERT's masked language modeling, has been adapted by Masked Autoencoders (MAE) [14]. These techniques fill in random missing patches and learn rich spatial and semantic features by training models to reconstruct missing pixels. This is especially important in the context of SSL methods, which decrease the necessity for high cost human annotated labels, thus enabling deep learning to be more scalable from one domain to another [8] [14].

C. Neural Architecture Search (NAS).

Neural Architecture Search is a way of automatically designing neural network architectures by searching through pre-defined architecture spaces to find the best architectures. To overcome this disadvantage, efficientNet, which was found by NAS, proposed a compound scaling which scales the depth, width, and

resolution uniformly. The result was that it could achieve higher accuracy than models designed by hand, with fewer parameters than the latter [9]. Lately, NAS methods include hardware-aware optimization, which makes sure that found architectures are efficient on target deployment hardware, such as mobile devices and edge computing hardware [15].

D. Foundation Models and Multimodal Learning

To fuse visual and textual understanding, a family of foundation models, like CLIP (Contrastive Language-Image Pre-training), is trained using 400 million image-text pairs [10]. CLIP achieves zero-shot transfer to a variety of visual classification tasks with minimal task-specific fine-tuning, showing that visual and linguistic supervision lead to more generalizable representations[26-27]. Few-shot learning has been explored in multimodal models such as "Flamingo," and it is believed that the future of image recognition will be based on multimodal intelligence and integration, not just vision.

IV. To implement and evaluate the RESENET50.

A. Dataset Preparation

The implementation is performed with CIFAR-10 dataset which consists of 10 classes of 60,000 images of color pictures of objects, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, and has been automatically divided into 50,000 training and 10,000 test images [17]. All pixel values are rescaled so they are in a range of [0.0, 1.0] by converting to float32 and dividing by 255. Better stability of the gradients, lower computational complexity, and faster convergence in backpropagation [18] are obtained in normalizing. The class labels are converted to one-hot encoded vectors to meet the requirement of categorical cross-entropy loss function.

B. Model Architecture

The implementation uses ResNet50 that has been pre-trained on ImageNet with include_top=False, which means that the top 1000-classifier is not included but instead the convolutional base is kept for feature extraction [6]. All base layers are frozen (layer.trainable=False) in order to keep pretrained representations in first training. The output tensor of ResNet50 is passed through a Flatten layer to flatten to a 1 dimensional vector of length 100,352, followed by a Dense(256, ReLU) layer to learn non-linear combinations of its features, a Dropout(0.5) layer for regularization, and a Dense(10, Softmax) classifier that computes probability distributions of the classes [6] [11].

C. Optimization and Training

The model has been compiled using the Adam optimizer with a learning rate of 0.0001, a combination of the RMSprop and momentum techniques for adaptive weight updates [19]. Adam computes the exponential moving average of gradient (first moment mt) and squared gradient (second moment vt) and updates the weights accordingly and applies bias to them. This helps the model to pass through smoother

loss regions without paying much attention to the complex regions, thus enhancing the convergence speed and stability [19].

The following callbacks are used to prevent overfitting, adjust LR if the validation loss has not decreased for a certain number of iterations, and save the best model automatically: EarlyStopping, ReduceLROnPlateau, ModelCheckpoint. Training curves exhibited consistent reduction of training and validation losses[20-22] as well as gradually increasing accuracy, with sharp loss reduction in the early epochs and levels off as the model converged [23-25].

D. Assessment and Outcomes

The model obtained the test loss value of 0.3658 and the test accuracy value of 87.28% on the CIFAR-10 test set. In prediction one can feed the test images into the network to get the logits and then apply softmax to transform those logits to meaningful probabilities. The final prediction of the model is the class with the largest softmax probability. Well-classed images have well-defined and high probability bars, and have high confidence, while misclassed images have flatter probability bars and have high uncertainty [6] [11].

Further improvement in results is recommended by fine tuning of the pretrained layers, more epochs and data augmentation. Further improvements to the model's generalization performance may be available with modern methods like progressive unfreezing of base layers [8] or self-supervised pretraining.

V. CRITICAL EVALUATION

The study of VGGNet, GoogLeNet, and ResNet shows that these networks are able to recognize images supervised, but have many differences in the gradient flow, efficiency of calculation, and the representation of features.

The stacked 3×3 convolutions in VGGNet were original & innovative for the time, and showed the strength of deep feature extraction [4]. Unfortunately, it has a very large number of parameters and lacks of any means of facing vanishing gradients, which makes it rather difficult to scale up. The computation and memory cost of the architecture is prohibitive for lighter and faster deployment scenarios.

The improved computational efficiency comes from the GoogLeNet's use of Inception modules that enable multi-scale feature capture using parallel convnets [5]. It's auxiliary classifiers allow for seamless gradient flow and structured learning of deeper networks. In the case of multiple branches and larger input dimensions, however, there can be a lot of overhead involved in the preprocessing, making it less suitable for smaller datasets.

ResNet is the best of the three models in maintaining the strength of the gradient during back propagation. Unlike VGGNet and GoogLeNet [6] that can only be trained with large networks, Identity skip connections can be used to train very deep networks while achieving higher accuracy with faster training time and less

computational complexity, and more reliable convergence. These theoretical benefits were also demonstrated in practice by using ResNet50 on CIFAR-10, with an accuracy of 87.28% obtained by transfer learning.

In the context of the larger AI field, ResNet's architectural breakthroughs are still incredibly relevant. ResNet backbones are commonly employed for feature extraction in Vision Transformers [7], object detection systems, and multimodal models [10]. In architectures such as DenseNet and UNet, the idea of skip connections has been generalized and has had a lasting impact [20]. However, recent advances like ViTs and foundation models are now starting to beat pure CNN architectures on large-scale benchmarks, indicating the best practices for current image recognition problems are hybrid models that utilise both convolutional extraction and attention mechanisms, and multimodal supervision [7, 10, 13].

VI. CONCLUSION

This paper has analyzed in detail three of the basic CNN architectures (ResNet, GoogLeNet, VGGNet) in terms of forward propagation, loss functions and backpropagation behavior. The models that are classical, ResNet shows the best performance, with the advantage of residual connections for solving the vanishing gradient problem and allowing training of significantly deeper networks. These theoretical benefits were confirmed by the ResNet50 design that was implemented in practice on CIFAR-10, using transfer learning to achieve an accuracy of 87.28% on the test set.

In addition to this comparative overview, we have looked into the emerging trends and advancements where modern Artificial Intelligence (AI) paradigms are continuing to expand and are complementing the CNN-inspired methods such as Vision Transformers, self-supervised learning, neural architecture search, and foundation models. The image recognition field is maturing and moving towards integrated systems with the inclusion of attention mechanisms, multimodal supervision and automatic architecture design. The principles behind ResNet, GoogLeNet and VGGNet remain fundamental blocks in a growing and complex AI stack as these technologies evolve.

REFERENCES

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

[2] GeeksforGeeks, “What is image recognition?” [Online]. Available: https://www.geeksforgeeks.org/computer-vision/what-is-image-recognition/

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

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

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–12.

[6] 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.

[7] 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

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

[9] 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

[10] 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.

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

[12] 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

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

[14] GeeksforGeeks, "Categorical cross-entropy in multi-class classification," [Online]. Available: https://www.geeksforgeeks.org/deep-learning/categorical-cross-entropy-in-multi-class-classification/

[15] 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

[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] 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

[21] 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.

[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] A. Krizhevsky, “CIFAR-10 dataset,” University of Toronto, 2023. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[24] K. Joshi, “Floating points in deep learning: Understanding the basics,” Medium, Jan. 2024.

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

[26] 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

[27] 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.