IJEMD-CSAI, Special Issue (2026) https://doi.org/10.54938/ijemdcsai.2026.04.2.638

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

Stable Backpropagation in Deep Image Recognition: A VGGNet-Centered Analysis with Modern AI Perspectives

Meng Chengshuo , Pan Junyu , Lei Kaisong , Deng Mile , Ryan Chia Chung Hern

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

Abstract

This paper explores the properties of back propagation and gradient flow of three basic CNN architectures – VGGNet, ResNet and Inception (GoogLeNet) – in image recognition. For each architecture, we explain the details of forward propagation, the calculation of categorical cross-entropy loss and the backward pass mechanisms. Training stability and transparent gradient monitoring is established in a VGGNet implementation on CIFAR-10. In addition to classical analysis, some modern AI paradigms such as Vision Transformers, self-supervised learning, neural architecture search, and foundation models are revolutionizing the field of image recognition, and are being introduced to augment the capabilities of CNN-based approaches.

Keywords : *VGGNet, backpropagation, gradient flow, convolutional neural networks, image recognition, Vision Transformers, self-supervised learning, CIFAR-10*

I. INTRODUCTION

Another crucial field of AI is image recognition, which empowers computers to decipher images and categorize them into meaningful classes [1]. The field has made progress from feature detection systems programmed manually to deep learning models which learn relevant features from vast amounts of data [2]. The most popular is the Convolutional Neural Network (CNN), inspired by biological visual processing, that are "stacked" to progressively learn simple edges up to complex shapes [3].

CNN is trained by using forward propagation to generate predictions, a loss function to quantify error, and back propagation to update weights by computing gradients [2]. The ability to capture the stability and efficiency of gradient flow during backprop is crucial for an effective learning ability of a model, especially for deep architectures where vanishing gradients are a major concern [4].

There are three representative architectures that deal with gradient flows differently: VGGNet [5] keeps gradient flow single, straight, and simple; ResNet [4] adds skip connections for direct gradient flow; and GoogLeNet [6] uses multi-branch Inception modules and auxiliary classifiers. Both approaches have their own pros and cons in terms of simplicity, efficiency, and scalability to depth[17-19].

AI advances have recently given rise to Vision Transformers (ViTs) equipped with self-attention mechanisms [7], self-supervised learning to minimise reliance on labelled data [8], and automated model architecture search [9] and foundation models such as CLIP for multimodal understanding [10]. In this paper, we analyze the gradient flow within the three classical architectures, implement a VGGNet on CIFAR-10 and place them in the modern AI context.

II. MODEL ANALYSIS

A. Forward Pass

In the forward propagation, the input image will go through convolutional layers that are used to apply the learnable filters, which will generate feature maps; then, the feature maps will undergo ReLU activation for nonlinearity and pooling for spatial reduction [3]. The operation performed at each convolutional layer is $Z(l) = W(l) * A(l-1) + b(l)$ where $A(l) = \max(0, Z(l))$ [1,2].

VGGNet processes the input data in a very sequential manner, consisting of three-step 3×3 convolutions with high interpretability [5]. ResNet introduces skip connections: $y = F(x, W) + x$, which helps to retain information throughout layers [4]. Inception uses multiple convolutions in parallel but uses different kernel sizes (1×1, 3×3, 5×5) and then concatenate outputs to create multi-scale feature extraction [6].

B. Loss Function

All three architectures rely on Categorical Cross-Entropy (CCE) 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-predicted probability [2, 11]. The loss is used as the optimization signal for backpropagation. GoogLeNet augments this with auxiliary losses at intermediate layers for enhanced gradient flow [6].

The backward pass and gradient flow are used.

Backpropagation applies the chain rule: $\partial L/\partial W(l) = (\partial L/\partial A(l)) \cdot (\partial A(l)/\partial Z(l)) \cdot (\partial Z(l)/\partial W(l))$ [2]. VGGNet has a single clear path for the propagation of its gradients, a feature known as a stacked structure. ReLU's

derivative(when active) is used to prevent gradient shrinkage, and the convolutional gradient formula ∇KL = conv(∇YL, rot180(X)) is used to ensure stable propagation [5].

The formula of ResNet introduces a "+1" term for the skip connections, so that the gradient of the loss function is never completely lost when the residual gradient is close to zero [4]: $\nabla xL = \nabla HL \cdot (\nabla xF(x) + 1)$. Gradients are distributed across 4 parallel branches in Inception: $\nabla xL = \sum \nabla YiL * Ki$, where auxiliary classifiers add some additional gradient injection points [6].

III. MODERN AI PARADIGMS

A. Vision Transformers

ViTs view images as sequences of patches and pass them through Transformer encoders that are based on self-attention [7]. They model global dependencies with a large scale of pretraining[20]. Hierarchical features in Swin Transformers are introduced by the shifted-window attention for practical performance [12].

B. Self-Supervised Learning

The contrastive learning (SimCLR, BYOL) and masked image modeling (MAE) learn robust visual representations without labels [8],[13]. These approaches not only decrease the amount of annotation but also enrich the representation for subsequent tasks[21-23].

D. Reinforcement Learning for Computer Vision

NAS automates the discovery of architecture. Superior Accuracy-Efficiency trade-offs are realized by efficient scaling with compound scaling, by EfficientNet [9]. Hardware-aware NAS guarantees deployability on resource constrained devices [14].

D. Foundation Models

Multimodal pretraining on image-text pairs has been shown to yield highly transferable representations in the context of zero-shot and few-shot visual tasks in CLIP [10] and Flamingo [15].

IV. THIS IS A SIMPLE IMPLEMENTATION OF VGG WITH ACCURACY ON CIFAR-10 DATASET.

A. Installation and System Design

Implementing a VGGNet on CIFAR-10 (32×32 color images, 10 classes) using TensorFlow/Keras. The network architecture is made up of three VGG blocks, each having 3x3 Conv-BatchNorm-ReLU layers followed by a max-pooling layer with filters of 64, 128 and 256 respectively. The dense flattening is

replaced by Global Average Pooling and softmax followed by dropout output. Throughout [5],[16] the following were used: He initialization and L2 regularization.

B. Training and Results

The training was done with SGD optimizer with learning rate of 0.01. Loss and accuracy curves exhibited smooth convergence indicating stability in training and validation metrics and showing close tracking, which implied a good flow of the gradient through the single-path architecture. The final accuracy obtained was 46.88% with accuracy of 48.46% on validation set, showing that the model is not severely overfitted, but learning was well balanced. The clear gradient trajectory allowed to easily observe the norms of the gradient and diagnose training behaviour [5], [11].

V. CRITICAL EVALUATION

VGGNet's single-path architecture offers the best visualisation of a gradient flow and the most explainable training dynamics, making it suitable for grasping the basics of backpropagation [5]. It's prone to vanishing gradients in very deep configurations, however, and also has a high number of parameters. ResNet's skip connections are the most effective at maintaining gradients across deep networks [4], whereas Inception's multi-branch architecture is efficient in terms of computation and allows for multiple paths of gradients [6].

In the era of AI, each of these architectures is used as a building block. Backbones are used in ResNet, which are widely used in Vision Transformers and object detection [7]. These fundamental concepts of the classical models also continue to guide the design of architectures for self-supervised learning, NAS and multi-modal systems [8, 9, 10].

VI. CONCLUSION

This paper showed that VGGNet's one-path design ensures stable and interpretable backpropagations for image recognition, and ResNet and GoogLeNet have better gradient management for deeper networks. These basic concepts are being expanded by emerging AI paradigms such as Vision Transformers, self-supervised learning, and foundation models, establishing a complex ecosystem where CNN-based architectures are still valuable components.

REFERENCES

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

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

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

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

[5] 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

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

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

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

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

[11] 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

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

[15] 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

[16] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training,” in Proc. ICML, 2015, pp. 448–456.

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

[18] 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

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

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

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

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, pp. 533–536, 1986.

[23] 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